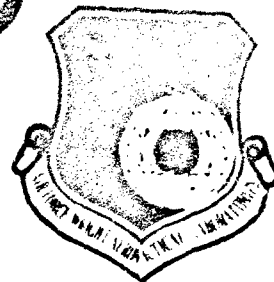


ADA114467

AFWAL-TR-81-4160  
VOLUME I



## EVALUATION OF NDE RELIABILITY CHARACTERIZATION

A. P. Berens  
P. W. Hovey

University of Dayton Research Institute

December 1981

Final Report for Period July 1980 - August 1981

20000728013

Approved for public release; distribution unlimited.

Reproduced From  
Best Available Copy

Materials Laboratory  
Air Force Wright Aeronautical Laboratories  
Air Force System Command  
Wright-Patterson Air Force Base, Ohio 45433

DTIC  
ELECTE  
S MAY 17 1982 D  
D

DTIC FILE COPY

02 05 11 049

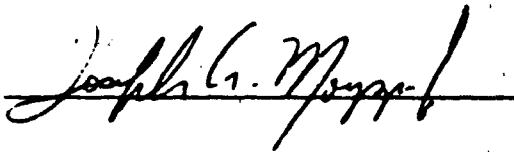
NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely related Government procurement operation, the United States Government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture use, or sell any patented invention that may in any way be related thereto.

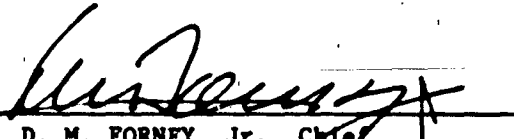
This report has been reviewed by the Office of Public Affairs (ASD/PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nations.

AFWAL-TR-81-4160, Volume II contains computer software, therefore distribution is limited in accordance with AFR 300-6 (DOD Dir. 4160.19, dated 5 April 1973). Non-DOD requests must include the statement of terms and conditions contained in Attachment 21 of AFR 300-6.

This technical report has been reviewed and is approved for publication.



FOR THE COMMANDER



D. M. FORNEY, Jr., Chief  
Nondestructive Evaluation Branch  
Metals and Ceramics Division

"If your address has changed, if you wish to be removed from our mailing list, or if the addressee is no longer employed by your organization please notify AFWAL/MLLP, W-PAFB, OH 45433 to help us maintain a current mailing list".

Copies of this report should not be returned unless return is required by security considerations, contractual obligations, or notice on a specific document.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFWAL-TR-81-4160, VOLUME I	2. GOVT ACCESSION NO. AD-A444 467	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) EVALUATION OF NDE RELIABILITY CHARACTERIZATION		5. TYPE OF REPORT & PERIOD COVERED FINAL July 1980 - August 1981
		6. PERFORMING ORG. REPORT NUMBER UDR-TR-81-113
7. AUTHOR(s) Alan P. Berens and Peter W. Hovey		8. CONTRACT OR GRANT NUMBER(s) F33615-80-C-5140
9. PERFORMING ORGANIZATION NAME AND ADDRESS University of Dayton Research Institute Dayton, Ohio 45469		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS P.E. 62102F 2418 05 22
11. CONTROLLING OFFICE NAME AND ADDRESS Materials Laboratory (AFWAL/MLLP) Air Force Wright Aeronautical Laboratories Wright-Patterson AFB, OH 45433		12. REPORT DATE December 1981
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		13. NUMBER OF PAGES 89
		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report)  Approved for public release, distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES AFWAL-TR-81-4160, Volume II contains computer software, therefore distribution is limited in accordance with AFR 300-6 (DOD Dir. 4160.19, dated 5 April 1973). Non-DOU requests must include the statement of terms and conditions contained in Attachment 21 of AFR 300-6.		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) non-destructive evaluation      crack detection reliability NDE reliability      inspection reliability NDE capability probability of crack detection		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) To characterize the uncertainty in non-destructive evaluation (NDE), a probability of crack detection (POD) as a function of crack length is postulated where POD is defined as the proportion of cracks of a given length that would be detected by the NDE technique when applied by inspectors to structural elements in a defined environment. This report: (1) presents a statistical framework for describing the uncertainty in the NDE determinations; and (2) evaluates various characterizations of NDE reliability. → over		

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

cont.)  
→ The data from a recent Air Force study on NDE reliability are used to estimate the parameters of the NDE model. For these representative capabilities NDE reliability experiments are simulated. Different NDE capability characterizations are computed for each simulated experiment and are statistically compared.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

# FOREWORD

This technical report summarizes the effort performed by the University of Dayton Research Institute under Materials Laboratory contract F33615-80-C-5140, "Evaluation of NDE Reliability Characterization," Task 24180522, in the area of characterizing the capabilities of non-destructive evaluation systems. The work was performed between July 1980 and August 1981. Dr. Joseph A. Moyzis of the Materials Laboratory was the Air Force Project Monitor and Dr. Alan P. Berens of the University of Dayton was Principal Investigator.

A complete description of the methods and results of the study are contained in Volume I. As part of the study, a computer program was written which simulates NDE experiments. Volume II comprises a user's manual and the listing for this simulation program.

Accession For	
NTIS GRA&I	<input checked="" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A	



## TABLE OF CONTENTS

<u>SECTION</u>		<u>PAGE</u>
1	INTRODUCTION	1
2	NDE RELIABILITY EXPERIMENTS	5
2.1	CATEGORY 1: NDE CAPABILITY AT ONE CRACK LENGTH	5
2.2	CATEGORY 2: ESTIMATION OF THE POD FUNCTION WITH ONE OBSERVATION PER CRACK	6
2.2.1	<u>Range Interval Method</u>	7
2.2.2	<u>Non-Overlapping Constant Sample Size Method</u>	9
2.2.3	<u>Overlapping Constant Sample Size Method</u>	9
2.2.4	<u>Optimized Probability Method</u>	11
2.3	CATEGORY 3: ESTIMATION OF THE POD FUNCTION WITH MULTIPLE OBSERVATIONS PER CRACK	14
3	A PROBABILISTIC FRAMEWORK FOR POD	17
3.1	THE BASIC FRAMEWORK	17
3.2	SELECTION OF POD(a) MODEL	21
3.3	ANALYSIS METHODS FOR NDE CHARACTERIZATION	24
3.3.1	<u>Regression Analysis</u>	25
3.3.2	<u>Maximum Likelihood Estimates</u>	30
4	EVALUATION OF NDE ANALYSIS METHODS	35
4.1	SIMULATED NDE CAPABILITY EXPERIMENTS	35
4.1.1	<u>Simulation Program</u>	35
4.1.2	<u>Experimental Conditions and Results</u>	38
4.2	EVALUATION OF RESULTS	39
4.2.1	<u>Comparison of Analysis Methods</u>	41
4.2.2	<u>Comparison of POD/CL</u>	54
4.2.3	<u>Comparison of Crack Size Distributions in the Specimens of an NDE Capability Experiment</u>	59
4.2.4	<u>Comparison of Scatter in Detection Probabilities at a Fixed Crack Length</u>	62

## TABLE OF CONTENTS (Concluded)

<u>SECTION</u>	<u>PAGE</u>
4.3 DISCUSSION	64
5 CONCLUSIONS	73
5.1 POD ANALYSIS FRAMEWORK AND REGRESSION MODEL SELECTION	73
5.2 RESULTS OF SIMULATION STUDIES	74
REFERENCES	77
APPENDIX A - "HAVE CRACKS WILL TRAVEL" DATA BASE	79

## LIST OF FIGURES

<u>FIGURE</u>		<u>PAGE</u>
1	Crack Growth-Life Curve at Second Inspection.	2
2	Example Confidence Limits from Range Interval Method.	8
3	Example Confidence Limits from Non-Overlapping Sixty-Point Method.	10
4	Example Confidence Limits From Overlapping Sixty-Point Method.	12
5	Example Confidence Limits From Optimized Probability Methods.	13
6	Example Confidence Limits From Reference 3 Analysis of Multiple Inspections at Each Crack.	15
7	Schematic Representation of Distribution of Detection Probabilities for Cracks of Fixed Length.	19
8	Example Application of Log Odds-Regression Analysis - Reference 3 Data.	28
9	Example 95% Confidence Limit Using Log Odds-Maximum Likelihood Analysis.	33
10	Flow Diagram of NDE Reliability Simulation.	36
11	POD Curves Used in the Simulation Study.	40
12	Comparison of Distributions of 90/95 Limits Produced by the Log Odds-Maximum Likelihood, Log Odds-Regression, and Optimized Probability Methods in Environment AET-Long Cracks.	45
13	Comparison of Distributions of 95/90 Limits Produced by the Log Odds-Maximum Likelihood, Log Odds-Regression, and Optimized Probability Methods in Environment AET-Long Cracks.	46
14	Comparison of Distributions of 90/95 Limits Produced by the Log Odds-Maximum Likelihood, Log Odds-Regression and Reference 3 Methods in Environment BET.	48



# LIST OF FIGURES (Continued)

<u>FIGURE</u>		<u>PAGE</u>
15	Comparison of Distributions of 95/90 Limits Produced by the Log Odds-Maximum Likelihood, Log Odds-Regression, and Reference 3 Methods in Environment BET.	49
16	Comparison of Distributions of 90/95 Limits Produced by the Log Odds-Maximum Likelihood, Log Odds-Regression, and Reference 3 Methods in Environment AUT.	50
17	Comparison of Distributions of 95/90 Limits Produced by the Log Odds-Maximum Likelihood, Log Odds-Regression, and Reference 3 Methods in Environment AUT.	51
18	Comparison of Distributions of 90/95 Limits Produced by the Log Odds-Maximum Likelihood, Log Odds-Regression and Reference 3 Methods in Environment AET.	52
19	Comparison of Distributions of 95/90 Limits for Log Odds-Maximum Likelihood, Log Odds-Regression and Reference 3 Methods in Environment AET.	53
20	Comparison of Distributions of 90/95 and 95/90 Limits Produced by the Log Odds-Regression Method in Environment BET.	58
21	Comparison of Specimen Crack Sizes in an NDE Capability Experiment - AET Environment, Log Odds-Regression Analysis.	60
22	Comparison of Specimen Crack Sizes in an NDE Capability Experiment - AET Environment, Log Odds-Maximum Likelihood Analysis.	61
23	Comparison of Degrees of Scatter in Detection Probabilities.	63
24	Effect of Scatter in Individual Crack Detection Probabilities on 90/95 Limits, AET-Long Cracks, Log Odds-Regression.	65
25	Effect of Scatter in Individual Detection Probabilities on 95/90 Limits, AET - Long Cracks, Log Odds-Regression.	66

# LIST OF FIGURES (Continued)

<u>FIGURE</u>		<u>PAGE</u>
26	Effect of Scatter in Individual Crack Detection Probabilities on 90/95 Limits, AET-Long Cracks, Log Odds-Maximum Likelihood.	67
27	Effect of Scatter in Individual Crack Detection Probabilities on 95/90 Limits, AET, Long Cracks, Log Odds-Maximum Likelihood.	68
28	Probability of Failing to Detect a Crack Greater Than Indicated Length.	71
A.1	Detection Percentages and Mean Log Odds Model for Cracks of AET Data Set - Have Cracks Will Travel Data Base.	83
A.2	Detection Percentages and Mean Log Odds Model for Cracks of AUA Data Set - Have Cracks Will Travel Data Base.	83
A.3	Detection Percentages and Mean Log Odds Model for Cracks of AUT Data Set - Have Cracks Will Travel Data Base.	84
A.4	Detection Percentages and Mean Log Odds Model for Cracks of BEO Data Set - Have Cracks Will Travel Data Base.	84
A.5	Detection Percentages and Mean Log Odds Model for Cracks of BET Data Set - Have Cracks Will Travel Data Base.	85
A.6	Detection Percentages and Mean Log Odds Model for Cracks of BRT Data Set - Have Cracks Will Travel Data Base.	85
A.7	Detection Percentages and Mean Log Odds Model for Cracks of CPT Data Set - Have Cracks Will Travel Data Base.	86
A.8	Detection Percentages and Mean Log Odds Model for Cracks of CUT Data Set - Have Cracks Will Travel Data Base.	86

LIST OF FIGURES (Concluded)

<u>FIGURE</u>		<u>PAGE</u>
A.9	Detection Percentages and Mean Log Odds Model for Cracks of EEA Data Set - Have Cracks Will Travel Data Base.	87
A.10	Detection Percentages and Mean Log Odds Model for Cracks of EEH Data Set - Have Cracks Will Travel Data Base.	87
A.11	Detection Percentages and Mean Log Odds Model for Cracks of FEA Data Set - Have Cracks Will Travel Data Base.	88
A.12	Detection Percentages and Mean Log Odds Model for Cracks of FEH Data Set - Have Cracks Will Travel Data Base.	88
A.13	Detection Percentages and Mean Log Odds Model for Cracks of FUT Data Set - Have Cracks Will Travel Data Base.	89

# LIST OF TABLES

<u>TABLE</u>		<u>PAGE</u>
1	EXAMPLE OF POD CALCULATION FROM DISTRIBUTION OF DETECTION PROBABILITIES	20
2	POTENTIAL POD FUNCTIONS	22
3	DEFINITION OF EXPERIMENTAL CONDITIONS USED IN SIMULATED NDE EXPERIMENTS	39
4	PRELIMINARY EVALUATION OF VARIOUS ESTIMATION PROCEDURES BASED ON 25-AET SIMULATED EXPERIMENTS	42
5	EVALUATION OF VARIOUS ESTIMATION PROCEDURES BASED ON 75 SIMULATED EXPERIMENTS - AET - LONG CRACKS	43
6	MEANS, STANDARD DEVIATIONS, AND COEFFICIENTS OF VARIATION FOR POD/CL LIMITS IN ENVIRONMENT BET ( $\alpha = -0.65$ $\beta = 0.88$ )	55
7	MEANS, STANDARD DEVIATIONS, AND COEFFICIENTS OF VARIATION FOR POD/CL LIMITS IN ENVIRONMENT AUT ( $\alpha = -3.9$ , $\beta = 1.8$ )	56
8	MEAN, STANDARD DEVIATIONS, AND COEFFICIENTS OF VARIATION FOR POD/CL LIMITS IN ENVIRONMENT AET ( $\alpha = -2.9$ , $\beta = 1.7$ )	57
A.1	"HAVE CRACKS WILL TRAVEL" DATA SETS	82

## SECTION 1

### INTRODUCTION

The United States Air Force approach to preventing structural fatigue failures is based on predictions of potential crack length as a function of flight time. For each critical area of the structure: (1) the length of the largest crack that could be present is ascertained; (2) the growth of this crack is predicted for the anticipated usage environment; (3) structural strength degradation due to crack growth is utilized to determine the safe time period during which the structure will not fail; and, (4) a repeat inspection is scheduled at half the flight time required for the crack to grow to critical size. This process is illustrated in Figure 1 which depicts projected potential crack growth in the anticipated usage environment through the second inspection. The initial crack length,  $a_0$ , is representative of manufacturing quality while the reset crack length at an inspection,  $a_{NDE}$ , is the longest crack that could pass undetected through the non-destructive evaluation (NDE) system. Since crack growth rates are highly dependent on crack length, the success of this procedure is greatly influenced by the correct choice of the initial and reset crack lengths.

A characteristic of all current non-destructive evaluation techniques is their inability to repeatedly produce correct indications when applied by various inspectors to flaws of the same "size". The ability and attitude of the operator, the geometry and material of the structure, the environment in which the inspection takes place, and the location, orientation, and size of the flaw all influence the chances of detection. However, since the structural maintenance actions are scheduled on the basis of potential crack length, the other factors are controlled (to the extent possible) or ignored and the resulting inspection uncertainty is characterized only in terms of crack length. A probability of crack detection (POD) for all cracks of a given length is postulated as the proportion of cracks that will be detected by an NDE system when applied by representative inspectors to a population of structural elements

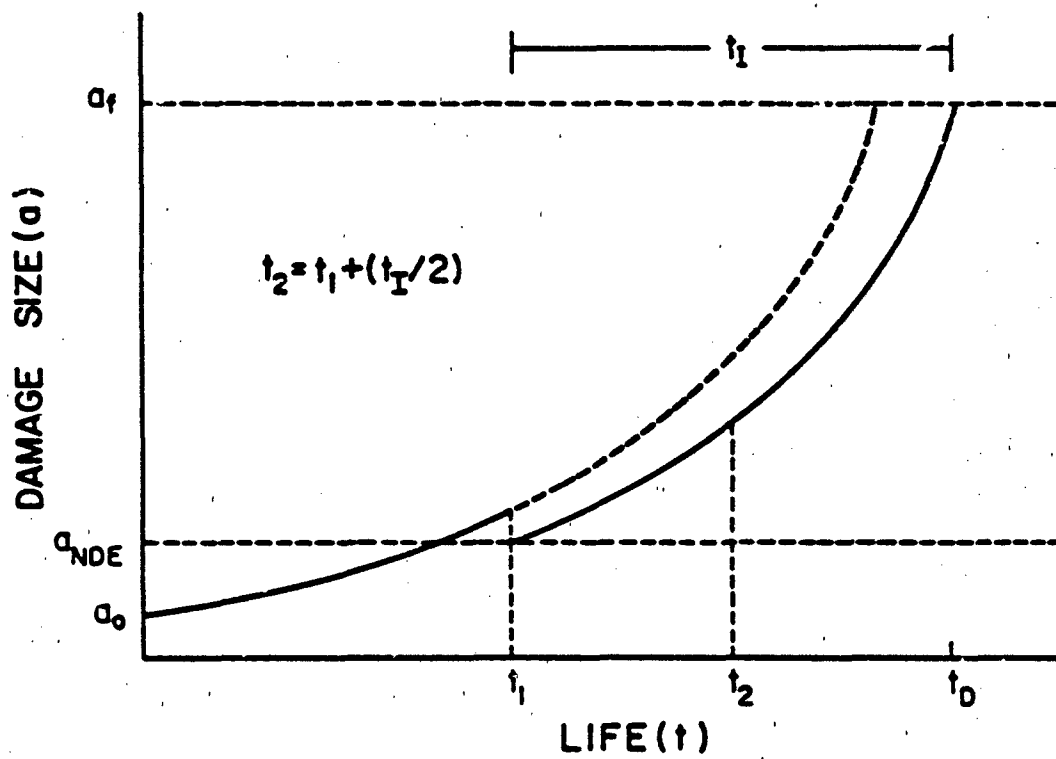


Figure 1. Crack Growth-Life Curve at Second Inspection.

in a defined environment. Thus, the capability of an NDE system is expressed in probabilistic terms and this characterization has two significant ramifications.

First, for a given NDE application, the true probability of detection as a function of crack length (or for a single crack length) will never be known exactly. The capability of an NDE system can only be demonstrated through an experiment in which representative structures with known crack lengths are inspected and the true POD is estimated by the observed percentage of correct positive indications. The estimated POD is subject to statistical variation that results from all uncontrolled factors that can lead to non-repeatable positive indications for cracks of a particular length. However, statistical methods (which depend on the experimental procedure) are available which yield confidence limits on the true probability. Protection against making a wrong decision on the basis of a set of non-typical results is provided by the confidence limits but an unknown element of risk will always be present.

Second, in the real world structural integrity problem, no inspection procedure will provide 100 percent assurance that all cracks greater than some useful length will be detected. The current capabilities and the uncertainty resulting from the NDE demonstration process at the short crack lengths of interest in aircraft applications dictate that the  $a_{NDE}$  value must be specified in terms of a high confidence that a high percentage of all cracks greater than  $a_{NDE}$  will be found. For example, MIL-A-83444 indicates that  $a_{NDE}$  is that crack length for which it can be shown there is 95 percent confidence that 90 percent of all cracks will be found. (Note that the chances of a crack longer than  $a_{NDE}$  passing undetected depends not only on the capability of the NDE system but also on the distribution of crack lengths that are present in the structural elements).

Due to the importance of the statistical characterization of NDE capabilities, this study was undertaken to: 1) evaluate and compare existing methods for determining  $a_{NDE}$  and the POD as a function of crack size; 2) to devise and evaluate different analysis

methods and models: and 3) to evaluate different combinations of POD and confidence limits as single number characterizations (a<sup>NDE</sup> values) of NDE systems. The essential tools used to achieve these objectives were the formulation of a probabilistic framework which recognizes that different cracks of the same length have different probabilities of detection and the use of this formulation to simulate NDE reliability demonstration experiments.

Section 2 presents a brief summary of the different types of NDE reliability demonstration experiments and the analysis methods that have been used for the resulting data. Section 3 presents the new probabilistic framework for modeling POD as well as two methods for estimating POD's within the framework. Section 4 describes the simulation process and contains the comparisons of different methods and different POD/confidence interval combinations. Conclusions and recommendations are contained in Section 5.

It should be noted that in general NDE reliability comprises two types of wrong indications: failure to give a positive indication in the presence of a crack and giving a positive indication when there is no crack (a false call). While it is recognized that both types of error are important, only the former is considered in this study. Due to the potential safety hazard, failure to find a crack is considered to be the more critical problem in the structural integrity application.



## SECTION 2

### NDE RELIABILITY EXPERIMENTS

There are three categories of experiments which have been used to evaluate the reliability of NDE systems: 1) demonstration of a capability at one crack length; 2) estimation of the POD function and confidence bounds through single inspections of cracks covering a range of lengths; and 3) estimation of the POD function and confidence bounds through multiple inspections of cracks covering a range of lengths. Analysis of data from category 1 and 2 experiments have generally been based only on binomial distribution theory. Category 3 data have been analyzed by regression analyses. Details of the three types of experiments and the analysis methods used to date are presented in References 1, 2, and 3. The following paragraphs summarize pertinent features of each current experiment/analysis combination. In the discussion, it is assumed that a representative population of inspectors are inspecting representative specimens in the environment of interest. Also, the question of the number of uncracked specimens submitted for inspection along with the cracked specimens is ignored even though this is recognized as an important experimental design condition. As noted earlier, false calls are being ignored in this study so only the cracked specimens enter the analysis.

#### 2.1 CATEGORY 1: NDE CAPABILITY AT ONE CRACK LENGTH

This category of experiments can be viewed as a method for demonstrating that an NDE system can detect at least a given percentage of cracks of a specified length with a specified confidence limit. For example, this approach could be used to show that there is 95 percent confidence that at least 90 percent of all cracks of length 0.25 in. will be found by a particular system. This category of experiments also serves as an introduction to the use of the binomial distribution for the analysis of NDE reliability data.

Out of a large number of specimens to be inspected, assume  $n$  contain a crack of length  $a$  and, at each inspection, there is a probability,  $p$ , of detecting the crack of length  $a$ . If  $r$  of the cracked specimens are detected,

$$\hat{p} = r/n \quad (1)$$

is an estimate of  $p$  and binomial distribution theory can be used to calculate lower confidence limits for the true, but unknown, value of  $p$ . The confidence limits depend on both  $r$  and  $n$  and are tabulated in many references, c.f., References 1 or 4.

To demonstrate a 90 percent POD at a 95 percent confidence level, a minimum of 29 specimens at the fixed crack length are required. If 28 out of 28 cracks are detected, the lower 95 percent confidence level for  $p$  is less than 0.9. Given the number of cracked specimens tested and the number of cracked specimens found, a lower confidence level for  $p$  can be constructed. If the lower confidence limit is sufficiently high (as defined prior to the experiment, e.g., 0.90) the demonstration is complete. If the lower confidence level for  $p$  is below the targeted value, the NDE system has failed the demonstration at that crack length.

This category of experiment provides information only at the crack length inspected and was not evaluated in this study. A complete discussion of the experiment and analysis method can be found in Reference 1.

## 2.2 CATEGORY 2: ESTIMATION OF THE POD FUNCTION WITH ONE OBSERVATION PER CRACK

In the second category of experiments many specimens covering a range of crack sizes are inspected once and the results are used to estimate the POD as a function of crack length with confidence limits. Different analysis methods have been used on the results of these experiments but they are all based on binomial distribution theory. Since, in general, there are not a large number of specimens with cracks of the same length, the specimens are grouped into intervals of crack length. It is assumed that all cracks

within a specified interval have approximately the same POD. The number of detections for the group is modeled by the binomial distribution, analyzed as described earlier, and the lower confidence bound for the POD is usually assigned to the crack length at the upper end of the interval.

The essential difference among the various methods of analyzing the data from this category of experiments has been in the method by which the crack length intervals are formed or combined. The following paragraphs present a brief summary of four analysis methods that have been used. References 1 and 2 present detailed discussions of these analyses. Reference 2 recommends the use of the method defined as the Optimized Probability Method (OPM) as described in Paragraph 2.2.4.

#### 2.2.1 Range Interval Method

In the Range Interval Method (RIM) crack length intervals are defined with equal lengths across the range of the data. Due to the somewhat random nature of the crack lengths in the specimens being inspected, the intervals constructed by this method will contain different numbers of flaws. The estimate of the POD function and its lower confidence limit can exhibit an apparent erratic behavior of the confidence limits resulted from small sample size. For example, Figure 2 presents RIM analyses of inspection results for eddy current inspections of etched fatigue cracks in 2219-T87 aluminum flat plates, Reference 2, page D-63. The individual data points are the percentage of cracks identified in each interval while the solid line segments connect the lower 95 percent confidence bounds for each POD estimate. The extremely erratic behavior of the confidence limits resulted from small sample sizes in some intervals even though all cracks greater than 3.6 mm were detected. For example, the very low confidence bound at about 6 mm resulted from that particular interval containing only one crack and, even though detected, the lower 95 percent confidence bound on  $p$  for a sample of size one is 0.05.

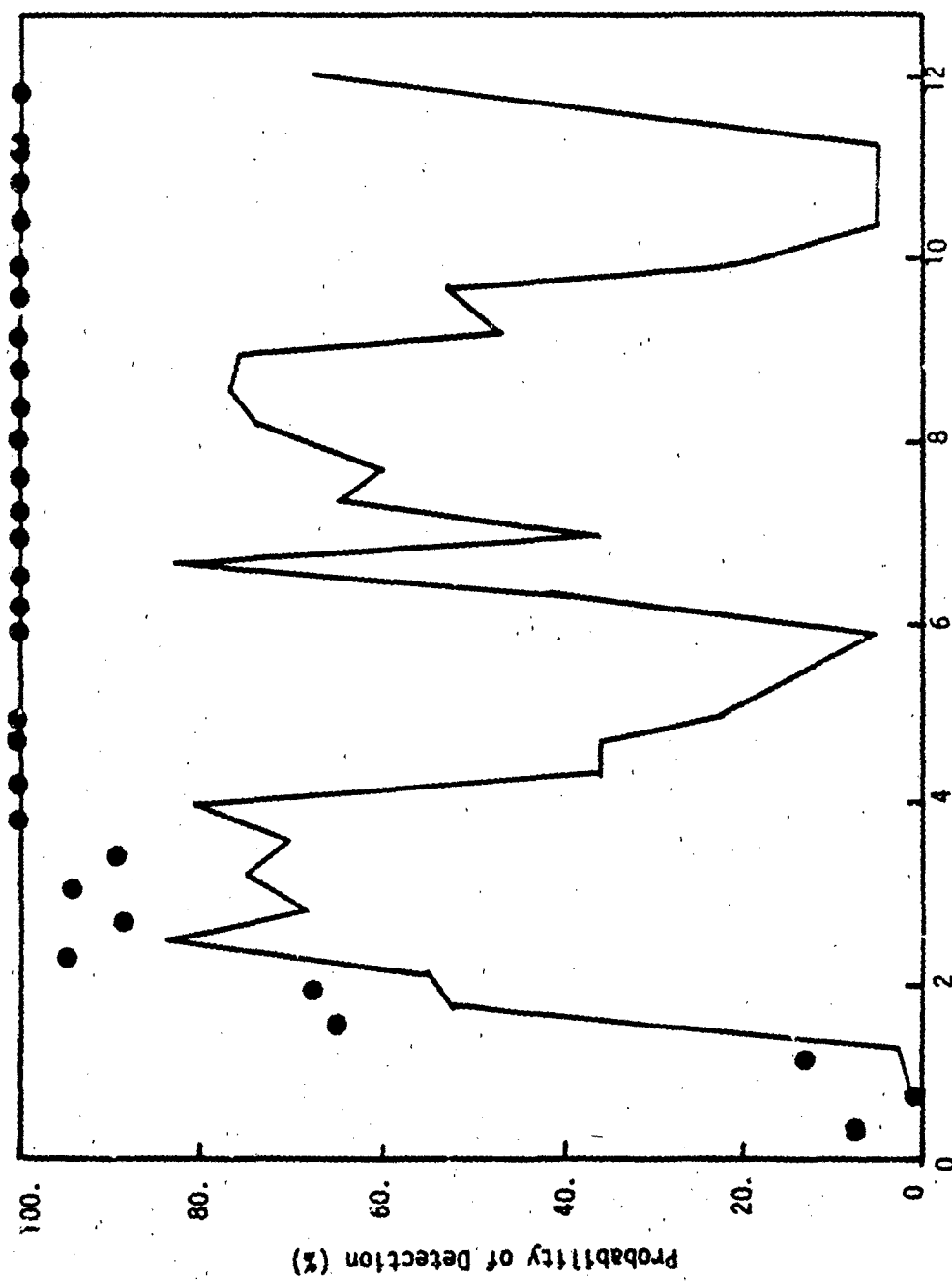


Figure 2. Example Confidence Limits from Range Interval Method.

### 2.2.2 Non-Overlapping Constant Sample Size Method

To avoid the problem resulting from the variable sample sizes, the length of the intervals can be varied so that each interval contains the same number of specimens with cracks to be detected. This method, the non-overlapping constant sample size method, is discussed in detail in Reference 1. If  $N$  represents the number of points in each interval, the intervals are constructed as follows. The longest  $N$  cracks form the first group. The second group contains the longest  $N$  cracks after the first group has been removed. The process is continued until all cracks are assigned to one of the intervals. Figure 3 shows the results of analyzing the data of Figure 2 by this method using a sample size of 60 cracks in each interval. As before, the data points represent the observed percentages of detections in each interval and the line segments connect the lower 95 percent confidence bounds. The data points and the lower confidence bounds are plotted at the longest crack size in each interval which introduces an unknown degree of conservatism in the NDE capability. Although in this example the lower confidence bound is monotonically increasing, other data sets could give rise to more erratic behavior; i.e., the monotonicity is a function of the data, not the analysis method. This method reduces the number of intervals for analysis which leaves longer gaps to be filled by interpolation. The number of points in each interval could be reduced but lowering the sample size in an interval widens the confidence bounds.

### 2.2.3 Overlapping Constant Sample Size Method

To obtain narrower intervals with a relatively large number of specimens in each crack length interval, the intervals are overlapped (Reference 2). Each interval contains the same number of specimens but the NDE indication for any particular specimen may be analyzed in more than one interval. The intervals are created by first grouping the largest  $N$  cracks. Of these, the smallest  $T$  percent is grouped with the next largest  $N-NT$  cracks. This process is continued until all cracks are assigned to at least one interval. As an example, in Reference 2, a 50 percent

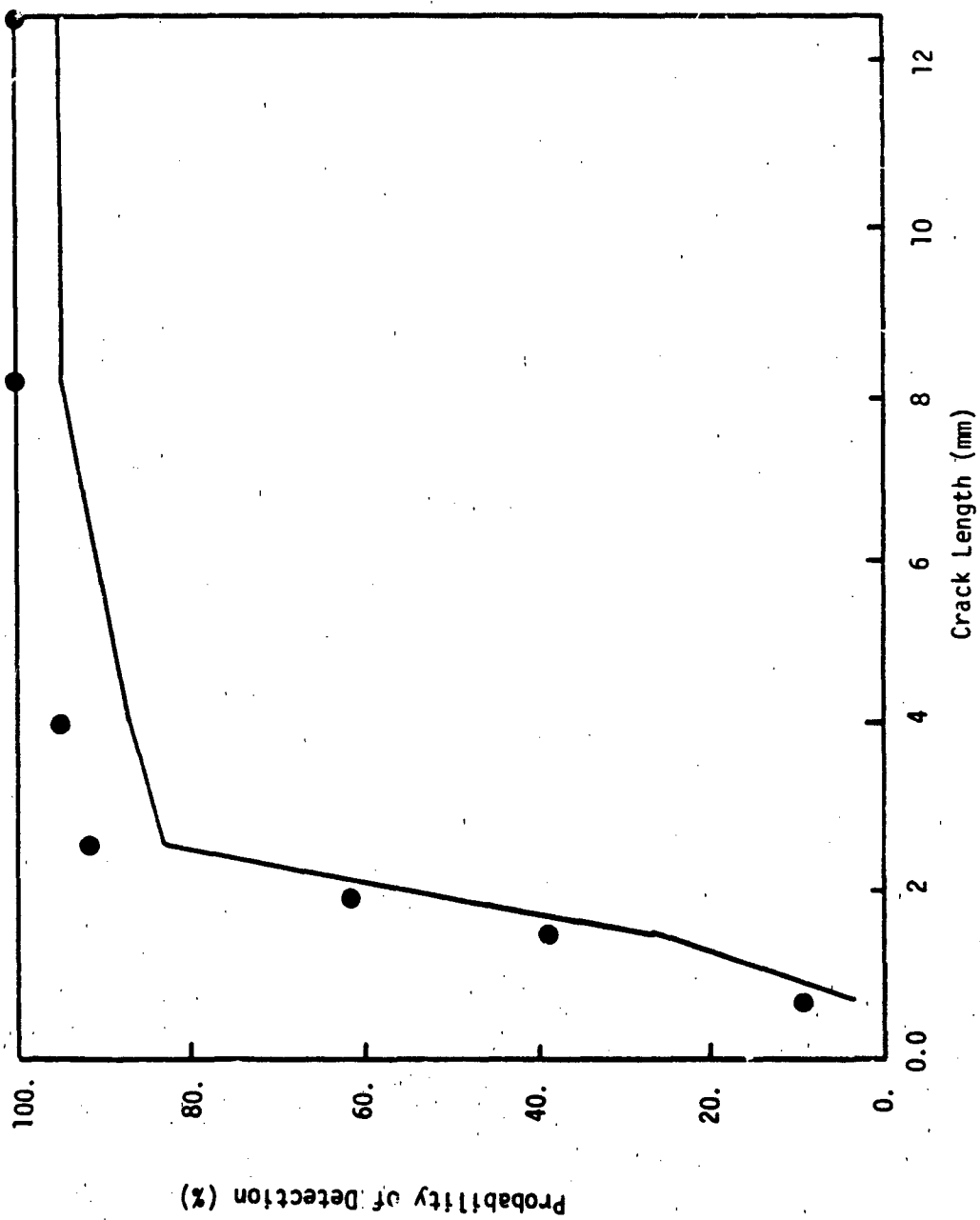


Figure 3. Example Confidence Limits from Non-Overlapping Sixty-Point Method.

overlap was used with a 60 point increment and was referred to as the overlapping sixty point method (OSPM). Figure 4 presents an example of the method for the data of Figures 2 and 3. Again, erratic behavior of the confidence bound can result and the bounds are plotted at the upper end of each interval with unknown conservatism. Further, using the results from a particular inspection in more than one interval results in deriving confidence limits from non-independent data sets with an unknown effect on the total measure of NDE capability.

#### 2.2.4 Optimized Probability Method

An algorithm for grouping experimental results to achieve the highest possible lower confidence bound from binomial distribution methods is presented in Reference 2. This method was named the optimized probability method. Initially, groupings are formed as in the range interval method giving rise to  $m$  intervals. For convenience, these intervals will be labeled 1 through  $m$  with the longest cracks in interval 1. The first OPM interval is selected as follows. The lower confidence bound on the POD is calculated for the data in interval 1, then for the data in intervals 1 and 2 together, then for intervals 1, 2 and 3 and so forth until a lower confidence bound for the POD is calculated for the whole data set (intervals 1 through  $m$  combined). The group of intervals that gives rise to the highest lower confidence bound is used as the first interval in the optimized probability method. Interval 1 data is then eliminated and the set of intervals 2 through  $m$  are analyzed in a similar manner to create the second OPM interval. This process continues until confidence bounds for all  $m$  intervals are found. Figure 5 shows the results of this procedure for the same data that was used in Figures 2 through 4.

Confidence bounds obtained by the optimized probability method are better behaved than those of the other interval. However, this behavior is obtained at the expense of unknown statistical validity of the POD function across all ranges of crack length.

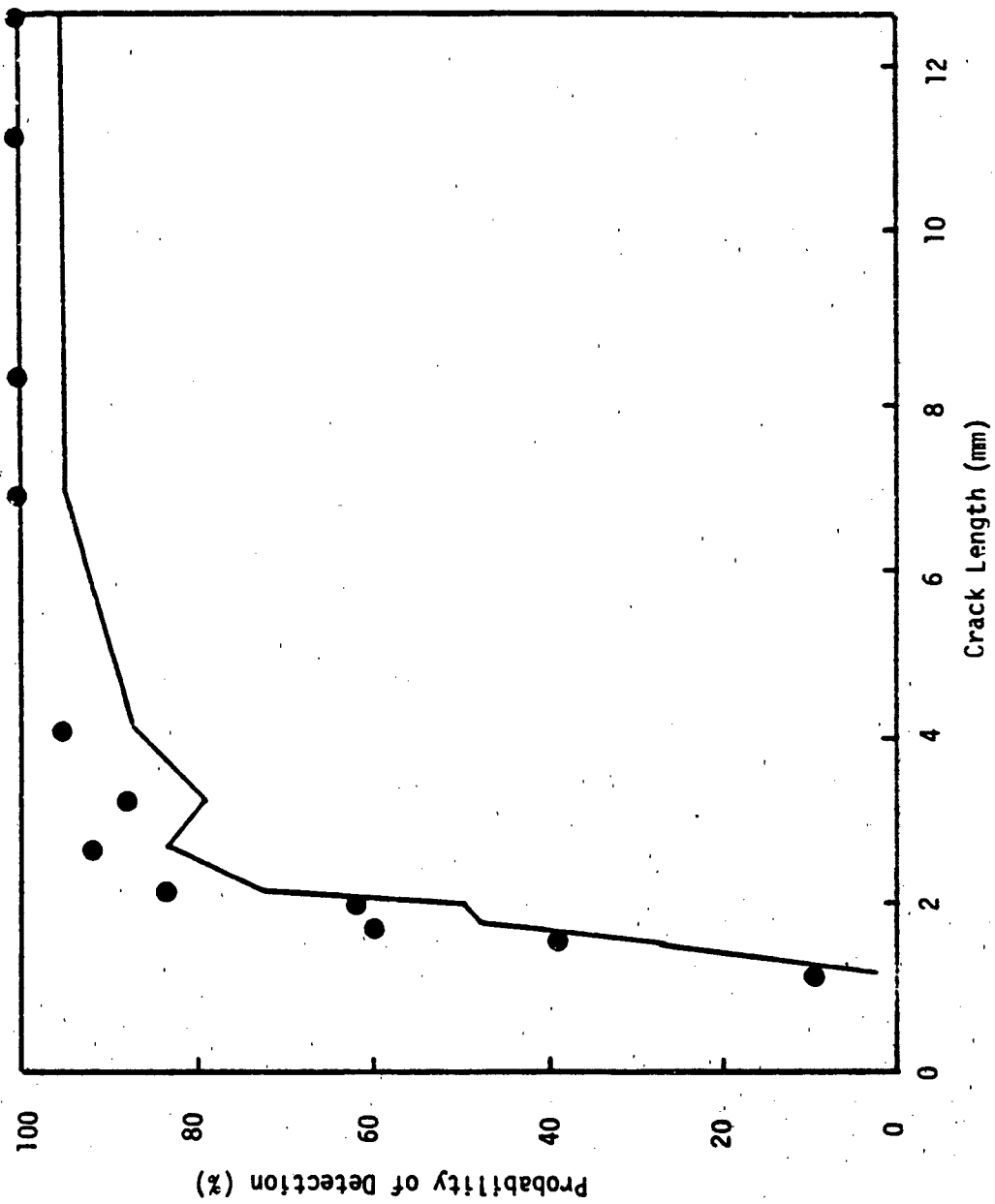


Figure 4. Example Confidence Limits From Overlapping Sixty-Point Method.



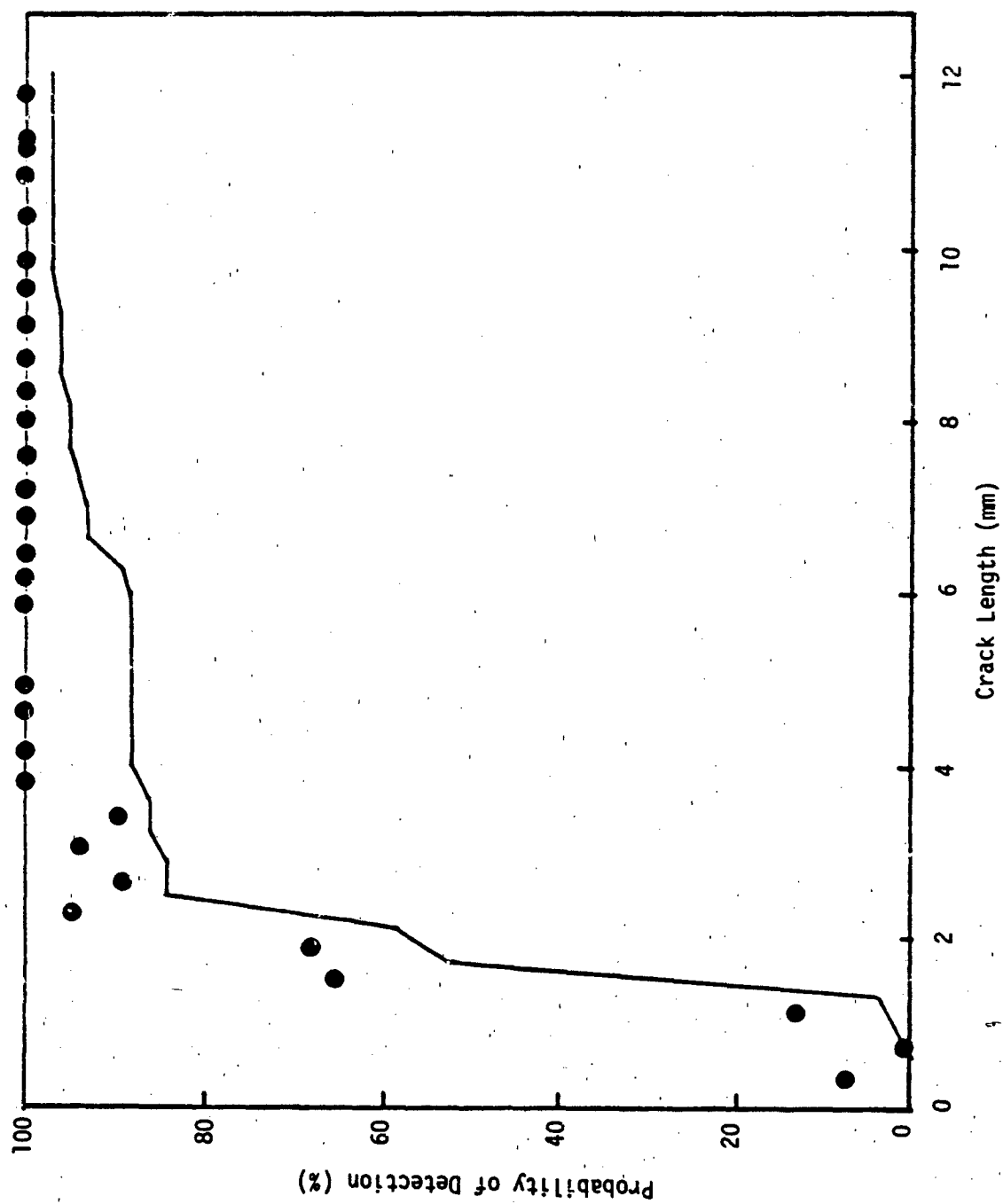


Figure 5. Example Confidence Limits From Optimized Probability Methods.

The overlapping of intervals requires inspection results for a particular crack to be analyzed more than once. Any crack that falls in an overlap area is used in calculating the confidence bounds for all the intervals involved. Thus, there is a correlation between confidence bounds that share data and the influence of this correlation on the POD as a function of crack length is unknown.

In general, the various "interval" methods have three major deficiencies. First, since they are based upon the binomial distribution, the confidence bounds are greatly influenced by the method for assigning cracks to intervals. The confidence bounds are as much influenced by the analysis method as they are the data. Second, the confidence bounds do not approach unity and, depending on the sizes of the cracks in the test specimen, may not reach the required 0.90 POD level. The entire experiment may fail to provide the desired result. Third, they provide limited inferences on the entire POD function if this function is required for further studies such as risk of failure analyses.

### 2.3 CATEGORY 3: ESTIMATION OF THE POD FUNCTION WITH MULTIPLE OBSERVATIONS PER CRACK

This category of experiment resulted from a large NDE reliability program performed for the Air Force (Reference 3 and Appendix A). Sections of retired aircraft and other specimens were transported to Air Force depots and inspected for cracks by representative personnel, using various NDE systems in a typical environment. At the completion of the travel phase of the program, the structures were destroyed to verify the existence and lengths of the cracks. This experiment has often been called the "Have-Cracks-Will-Travel Program" and the data from the program is often referred to as the "Have-Cracks" data.

This method of collecting data yields an estimate of a detection probability for each individual crack. For example, Figure 6 from data of Reference 3, displays the inspection results of individual cracks emanating from fastener holes in a skin and stringer wing assembly (Sample A) as inspected by eddy current

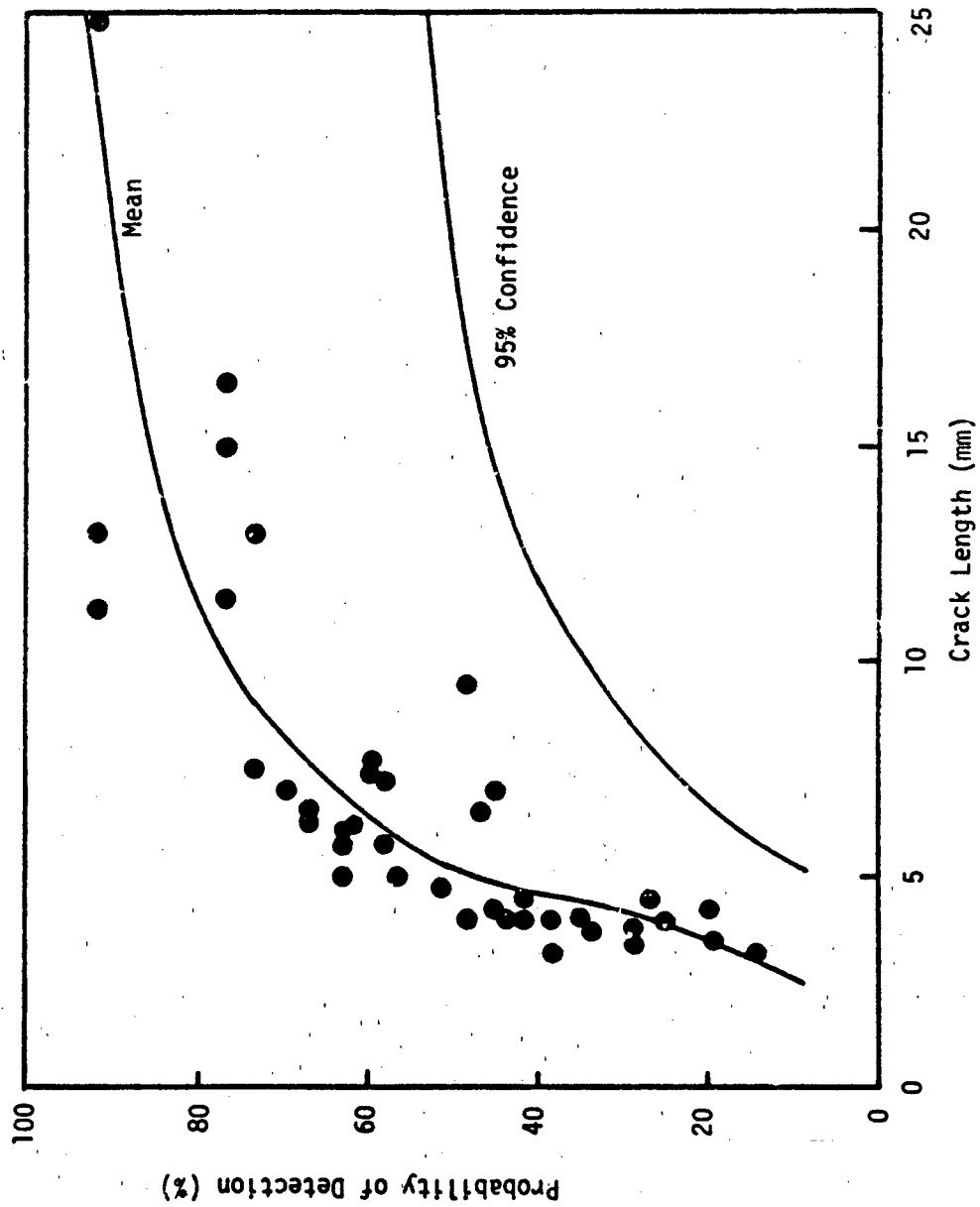


Figure 6. Example Confidence Limits From Reference 3 Analysis of Multiple Inspections at Each Crack.

surface scans. Each data point in the figure represents the proportion of times the crack was found when subjected to 60 independent inspections by different inspectors. These data clearly illustrate that not all cracks of the same length have the same detection probability.

To analyze the data collected from this category of NDE experiment, a regression analysis is performed in which a model curve is fit to data points (as represented by Figure 6) and a lower confidence limit is placed on the regression equation. In Reference 3, the model selected as providing the best fit is given by

$$\text{POD}(a) = \exp[-\alpha a^{(1-\beta)}] \quad (2)$$

where the parameters are estimated by a linear regression on transformations of the crack lengths and observed probabilities of detection. The curve labeled "mean" in Figure 6 illustrates the fit of the model to the data points.

In Reference 3, the viewpoint was taken that the POD at a particular crack length was a low percentile of the distribution of detection probabilities at the crack length. To calculate a lower confidence limit on the POD, a confidence bound on the population of detection probabilities was used. Such a bound is also shown in Figure 6 as the "95% Confidence" curve. In this particular example, the "95% Confidence" curve is considerably below all of the data points. It will be shown later that the POD is actually the mean of the detection probability distribution. Hence, the POD confidence limits should be placed on the mean regression line not on the total population of detection probabilities.

### SECTION 3

#### A PROBABILISTIC FRAMEWORK FOR POD

Prior to the "Have Cracks" program described in Reference 3, the results of all inspections of details with cracks of approximately equal length were treated as completely random samples from Bernoulli trials with a constant probability of successful detection. This model is an entirely correct approach for completely independent inspections. In the "Have Cracks" program, the same details were inspected by a large number of inspectors. The repeated observations on the same details required a different analysis model which separately handled the different detection probabilities for different details. A regression approach was used to analyze these data but no rationale other than goodness of fit was presented for the approach. In the following, a probabilistic framework is presented in which it is shown that the POD is the average of all detection probabilities of all cracks of the same length. Thus, the regression equation provides an estimate of the POD function, but a regression model must be determined. The "Have Cracks" data were used to compare and select an acceptable model from seven candidates. For the best model, methods of estimating the parameters are also discussed.

#### 3.1 THE BASIC FRAMEWORK

In the data of the "Have Cracks" program, there are many examples of two cracks of approximately equal length having significantly different probabilities of being detected (c.f. Figure 6 and Appendix A). These results demonstrate that the POD is influenced by many factors and is only correlated with crack length. To model the POD, assume that there is a distribution of detection probabilities at each crack length where the scatter in this distribution is caused by the non-reproducibility of all factors other than crack length. Examples of such factors are differences in detectability due to operators, environments, and crack orientation, geometry or location. The probability of detection of a crack selected at random from all cracks of a fixed length can be calculated as follows.

Let  $f_a(p)$  represent the density function of the detection probabilities for the population of details which have a crack length of  $a$ . Figure 7 presents a schematic representation of this distribution. The probability of selecting a crack whose detection probability is between  $p$  and  $p+dp$  is  $f_a(p) dp$  by definition of a density function. The probability that this crack will be detected is  $p$ . The conditional probability that the detection probability is  $p$  and that a positive indication will be given is  $p f_a(p) dp$ . To find the unconditional probability that a randomly selected crack of length  $a$  will be detected,  $POD(a)$ , the conditional probabilities are summed over the range of detection probabilities. Therefore,

$$POD(a) = \int_0^1 p f_a(p) dp \quad (3)$$

In addition to calculating  $POD(a)$ , equation 3 is also the formula for the mean of the distribution of detection probabilities at crack length  $a$ .

The following simple example illustrates this property. Assume at some crack length there are only five types of cracks and for each crack type there is a constant detection probability as given in Table 1. Assume, further, that the five crack types are equally represented in the structural details of interest. Under these conditions, the average of the detection probabilities,  $POD(a)$ , is 0.80. If 1,000 inspections are performed on details chosen at random, 200 of each type would be expected in the total sample. Of the 200 cracks of Type 1,  $(0.65) (200) = 130$  would be expected to be detected and similarly for the other four crack types as indicated by the calculations of Table 1. The total number of cracks that would be expected to be found is 800 or 80 percent of the total number inspected. Thus, the percentage found would be expected to be the  $POD$  as calculated by equation (3).

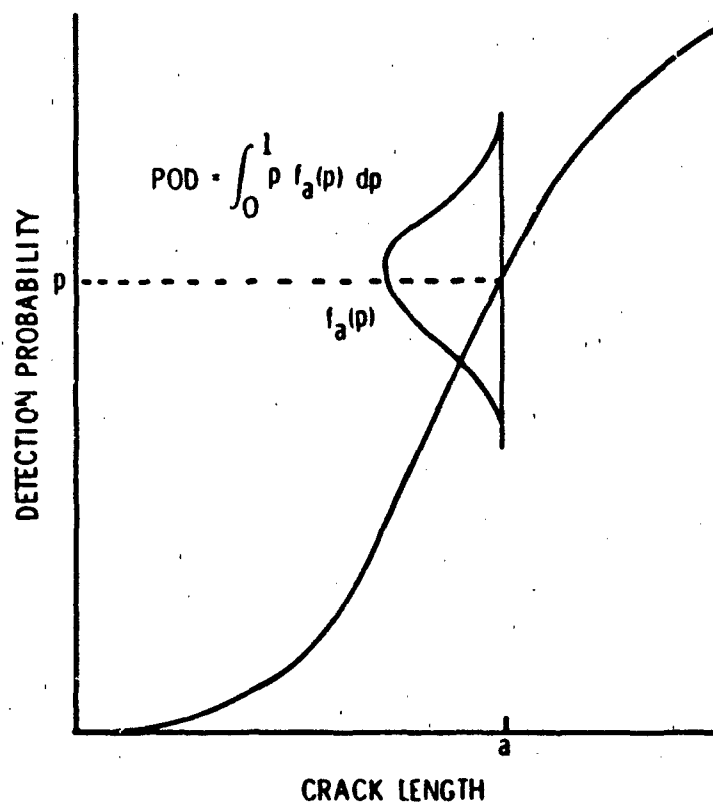


Figure 7. Schematic Representation of Distribution of Detection Probabilities for Cracks of Fixed Length.

TABLE 1  
EXAMPLE OF POD CALCULATION FROM  
DISTRIBUTION OF DETECTION PROBABILITIES

<u>Crack Type</u>	<u>p</u>	<u>Expected # Times in 1000 Trials</u>	<u>Expected # of Detections</u>
1	0.65	200	130
2	0.75	200	150
3	0.80	200	160
4	0.85	200	170
5	0.95	200	190
<hr/>			
TOTAL	$\bar{p} = 0.80$	1,000	800 = 80% of trials



As shown by equation (3), the true POD function is the curve through the averages (means) of the individual density function of detection probabilities. Such a curve is also the traditional regression equation of POD as a function of crack length. Therefore, regression analysis techniques can be used to estimate  $POD(a)$  when individual estimates of the detection probabilities are available. Confidence limits on the true POD, however, would be calculated from the confidence limits for the average of the predictions, not for the individual detection probability estimates.

### 3.2 SELECTION OF $POD(a)$ MODEL

Although  $POD(a)$  has been shown to be the curve through the mean of detection probabilities, a functional model for this curve must be determined or assumed before the characterization can be used. Since the "Have Cracks" data are representative of field inspection capability for selected structures and inspection methods and are the largest data set for which detection probabilities have been estimated for many cracks from many inspectors, they were selected for a study to determine an acceptable model for the  $POD(a)$  function. See Appendix A for a discussion of these data. Three criteria were established for the definition of "acceptable": (1) goodness of fit, (2) normality of deviations from fit, and (3) equality of variance of deviations from fit for all crack lengths. The latter two criteria are necessary statistical assumptions for the validity of confidence limits derived from regression analyses.

Seven functional forms were investigated as listed in Table 2. The Lockheed model was derived during the original analysis of the "Have Cracks" data. The Weibull model was selected since it is a generally accepted model and is a variation of the Lockheed model. The other five models were selected because they have been found useful in analogous problems in the field of bioassay (References 5 and 6). In particular, the transformations have been derived to yield the properties of normality and equality of variance (the second and third criteria). The probit and log probit models used the normal distribution to transform the observed detection

TABLE 2  
POTENTIAL POD FUNCTIONS

<u>Name</u>	<u>Functional Form</u>	<u>Transformation</u>
Lockheed	$P(a) = e^{-\alpha(a)^{1-\beta}}$	$Y = \ln \left( \frac{-\ln(p)}{a} \right), X = -\ln(a)$
Weibull	$P(a) = 1 - e^{-\alpha(a)^\beta}$	$Y = \ln(-\ln(1-p)), X = \ln(a)$
Probit	$P(a) = \Phi(\alpha + \beta a)^*$	$Y = \text{PROBIT}(p), X = a$
Log Probit	$P(a) = \Phi(\alpha + \beta \ln(a))^*$	$Y = \text{PROBIT}(p), X = \ln(a)$
Log Odds -linear scale	$P(a) = \frac{e^{\alpha+\beta a}}{1 + e^{\alpha+\beta a}}$	$Y = \ln(p/(1-p)), X = a$
Log Odds -log scale	$P(a) = \frac{\alpha a^\beta}{1 + \alpha a^\beta}$	$Y = \ln(p/(1-p)), X = \ln a$
Arcsine	$P(a) = \sin^2(\alpha + \beta a)$ $0 \leq a \leq \frac{\pi - 2\alpha}{\beta}$	$Y = \arcsine(\sqrt{p}), X = a$

---


$$* \quad \Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy$$

probabilities. A probit is the value obtained by applying the inverse normal distribution function. The log odds or logistic model uses the logistic function to transform the detection probabilities. The logistic function is an approximation to the normal distribution function and, in addition, equalizes variances. The arcsine model was included because of its use in bioassay even though there could be practical difficulties in its use as a POD(a) model.

Regression analyses were used to fit all seven models to the "Have Cracks" data. The "Have Cracks" data comprise 13 data sets with a data set being defined in terms of NDE method and type of structure. Each model was fit to all 13 data sets. The detection probabilities,  $p_i$ , and the crack lengths,  $a_i$ , for each crack were transformed to  $Y_i$  and  $X_i$  in accordance with the transformations of Table 2. The transformed X and Y variables were then used in a linear regression analysis of the form

$$Y_i = A + BX_i + e_i \quad (4)$$

For all seven models, B is the estimate of  $\beta$  and, depending on the model, either A or  $\exp(A)$  is the estimate of  $\alpha$ . The deviations of the transformed observations from the regression equation,  $e_i$ , were analyzed to test the applicability of each model with respect to the three acceptability criteria.

The probit, log odds-linear and arcsine models were based on models which did not transform the crack length scale. None of these models provided adequate goodness of fit in the sense that their patterns of deviations were not randomly distributed about the model over the entire crack length range (i.e., they were inconsistent with the linear model of equation (4)). These models were eliminated on this basis. The other four models generally provided an adequate fit to the observed data.

The Bartlett's test and the Shapiro-Wilks W test were used to evaluate the equality of variance and normality, respectively, of the deviations from the regression equations (Reference 7). The variance of the deviation from the log odds-log scale and log

normal models was constant (within random error) across the range of crack lengths in all data sets. Equality of variance was rejected for some crack length/data set combinations for the Weibull and Reference 3 models.

The log odds-log scale model was consistent with the assumption of normality of deviations in all but two data sets. Consistency was defined as the failure to reject the hypothesis at the 0.1 level of significance. In the two cases for which normality was rejected for the log odds-log scale model, the probability of the type I error was between 0.05 and 0.10. None of the other models performed as consistently. Therefore, it was concluded that the log odds-log scale model provided an adequate fit to the POD(a) function for the "Have Cracks" data (see Appendix A). This model, hereafter referred to as the log odds model, was adopted for the simulation analysis study.

### 3.3 ANALYSIS METHODS FOR NDE CHARACTERIZATION

The analysis framework of the previous section demonstrates that the POD function is the mean of all detection probabilities at each crack length. This framework indicates that NDE capabilities can be characterized by regression analyses if a functional form is assumed for POD(a). In particular the equation

$$POD(a) = \frac{\exp(\alpha + \beta \ln a)}{1 + \exp(\alpha + \beta \ln a)} \quad (5)$$

was shown to fit the "Have Cracks" data and may be an acceptable model for other NDE applications.

The following paragraphs present methods for estimating the POD and confidence limits on the POD. The regression analyses of this section are based on any POD model which can be transformed to a linear equation in detection probability and crack length. In addition, a maximum likelihood estimation method is presented but the equations are strictly dependent on the model of equation (5). A different model would require different equations for the estimates of the parameters.

### 3.3.1 Regression Analysis

Equation (3) indicates that the parameters of the POD(a) function can be estimated through a regression analysis. Further, it has been shown that for the "Have Cracks" data, equation (5) is a model of POD(a) for which the assumptions of normality and equality of variance of the residuals, is acceptable. Therefore, equation (5) can be fit to the data and standard statistical regression methods can be used to find a lower confidence bound on the true POD(a).

If the data are from the category 3 type of experiment in which the detection probability for each crack is estimated through multiple inspections, the  $(a_i, p_i)$  data pairs can be input directly into the transformations required by the analysis. If the data are from a category 2 type of experiment in which each of many cracks is inspected once, the cracks will have to be grouped into intervals of crack length and the proportion of detections assigned to a single crack length representative of the interval. Since the detection probability for an interval is more likely to be representative of the middle of the interval than the endpoints, the middle is recommended. However, using the upper endpoint as a conservative measure could be employed, but the probability level of the resulting confidence bound would not necessarily be correct.

Given the  $n$  pairs of  $(a_i, p_i)$  data points to be fit by the regression analysis, the appropriate transformations are performed. For the model of equation (5), the transformations are defined by

$$Y_i = \log (p_i / (1 - p_i)) \quad (6)$$

and

$$X_i = \log (a_i) \quad (7)$$

The variables X and Y are then used in a linear regression analysis resulting in estimates A and B for  $\alpha$  and  $\beta$ , respectively. The formulas for A and B are

$$B = \frac{\sum_{i=1}^n X_i Y_i - \frac{\sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n}}{\sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n}} \quad (8)$$

$$A = \bar{Y} - B \bar{X} \quad (9)$$

where  $\bar{Y}$  and  $\bar{X}$  are given by

$$\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}, \quad \bar{X} = \frac{\sum_{i=1}^n X_i}{n} \quad (10)$$

The formula for a lower confidence bound on the mean  $\mu_{Y_L}$  for a given value is

$$\mu_{Y_L} = A + B X - t_{(n-2), \gamma} S_{Y.X} \sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{SSX}} \quad (11)$$

where

$\gamma$  is the confidence coefficient

$t_{(n-2), \gamma}$  is the  $\gamma$ th percentile of a t distribution with n-2 degrees of freedom

$$S_{Y.X} = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (Y_i - A - B X_i)^2} \quad (12)$$

$$SSX = \sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n} \quad (13)$$

Since the log odds transformation is monotonic, the reverse transformation of the confidence bound on  $Y(a)$  will be the confidence bound on  $POD(a)$ . Although somewhat complicated algebraically, an inverse to the confidence bound function can be derived explicitly. This formula can be used to calculate the single point NDE capability characterizations expressed in terms of a  $POD$ /confidence level.

Figure 8 presents the results of the regression analysis of the "Have Cracks" data of Figure 6 using the log odds model of equation (5). The solid line represents the lower 95 percent confidence bound on the true  $POD(a)$ . The results of this analysis should be contrasted with those of Figure 6 which were calculated using the analysis methods of Reference 3. The mean or best estimate curves are essentially the same in both figures as the functional forms of both the Lockheed and log odds models fit the data reasonably well. The significant difference in the 95% confidence curves results primarily from the method of calculating the confidence limits from the mean curve and, consequently, different interpretations of the 95% confidence curves are required.

As shown in subsection 3.1, the probability of detection of a randomly selected crack of a fixed length is equal to the mean of all detection probabilities for cracks of that length. The mean curves of Figures 6 and 8 are both regression estimates of this average as a function of crack length. The confidence limit curve of Figure 8 was calculated such that there is 95% confidence that the mean detection probability ( $POD$ ) lies above the curve for any single crack length. The confidence limit curve of Figure 6 was calculated such that a randomly selected crack at a fixed length will have a 95% chance of having a detection probability greater than the indicated value, i.e., that 95% of the individual data points would lie above the confidence bound. Since confidence bounds on individual elements of a population are always wider than equivalent bounds on the mean, the confidence limit curve of Figure 6 must be lower than that of Figure 8. The difference between the confidence bounds is determined by the amount

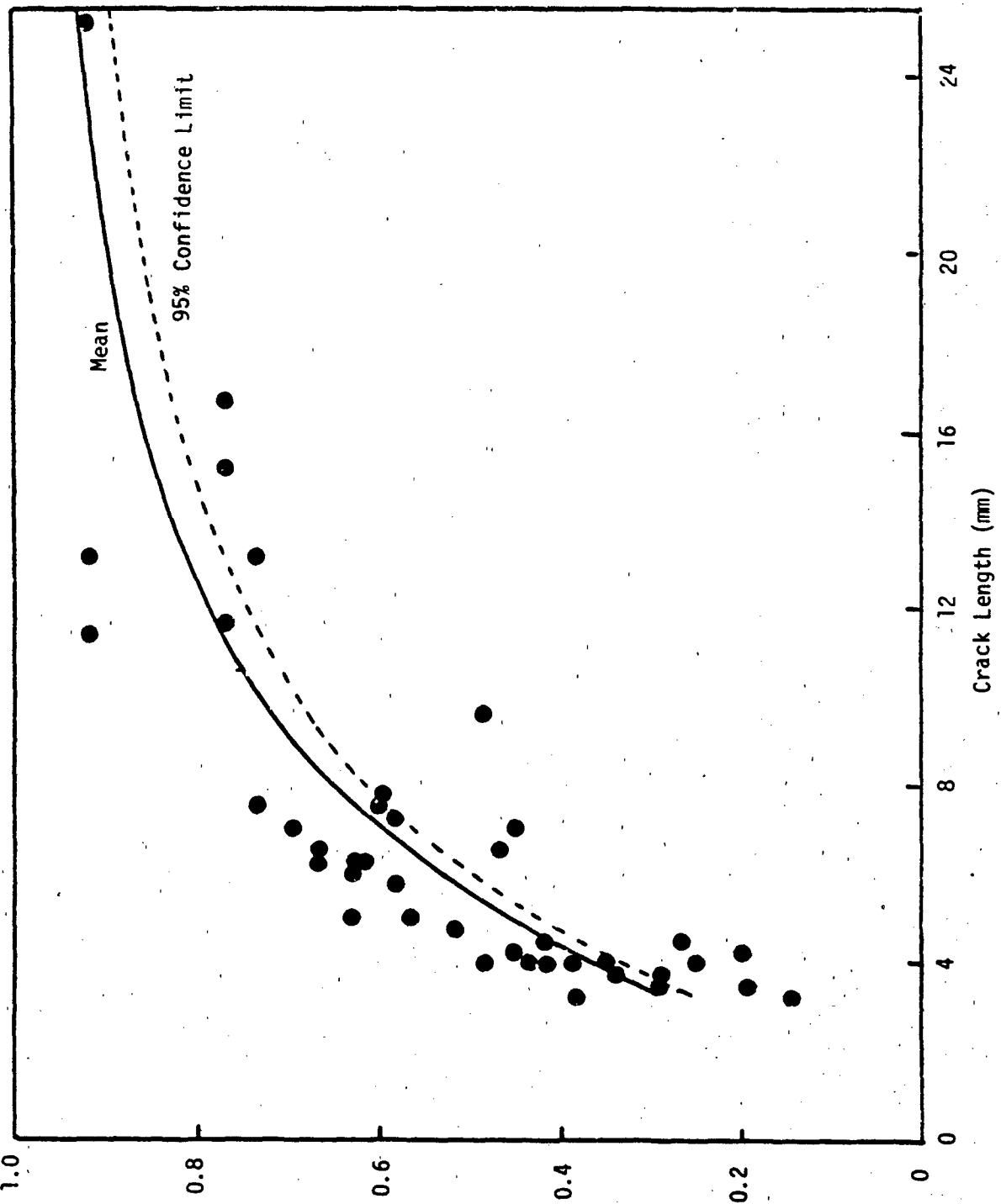


Figure 8. Example Application of Log Odds-Regression Analysis - Reference 3 Data.



of the scatter of the individual data points about the mean. While both of these analyses characterize NDE capability, the confidence bound of Figure 8 is the regression equivalent of the classical confidence bounds using binomial methods. The approach of Figure 6 does not yield confidence limits on the POD and, if used, another name should be given to this type of more conservative characterization.

A problem in the use of regression analysis arises when the observed proportion of detected cracks at a crack length is zero or one. In either of these cases, the most useful transformations can be undefined. To circumvent this problem, there are several alternatives. In Reference 3, the values of 0.01 and 0.999 were substituted for 0 and 1, respectively. However, the regression results are sensitive to the arbitrarily defined values. A more acceptable solution is to use a different estimator for the detection probability.

The usual (maximum likelihood) estimator for the detection probability is taken as

$$\hat{p} = i/n \quad (14)$$

where  $i$  is the number of detections and  $n$  is the number of specimens with the crack of the fixed length. Other estimates of the proportion which have acceptable statistical properties are the mean estimate

$$\bar{p} = i/n+1 \quad (15)$$

and the median estimate:

$$\tilde{p} = \frac{i-0.3}{n+0.4} \quad (16)$$

In this study, equivalent mean estimates were used if  $\hat{p}_i=0$  or  $\hat{p}_i=1$ . If a crack was detected by all inspectors ( $\hat{p}_i=1$ ), the detection probability was set equal to  $n/n+1$ . If a crack was never detected ( $\hat{p}_i=0$ ), the detection probability was set equal to  $1/(n+1)$  which is the equivalent result that would be obtained if the definitions of success and failure were reversed. (If the median estimates,  $\tilde{p}$ , were used when  $i=0$  or  $n$ , a similar equivalent

formula for the  $i=0$  case would be required.) These substitutions were judged to have no significant effect on the results of this study.

### 3.3.2 Maximum Likelihood Estimates

Given the POD(a) model of equation (5), an entirely different method for estimating the parameters uses the principal of maximum likelihood. In this type of estimation the parameter estimates are the values which maximize the probability of obtaining the observed data. The maximum likelihood estimates do not require grouping of data when the experiment is the single inspection per crack of category 2. Instead, they are based directly on the observed outcomes of 0 for a non-detection and 1 for a detection. This paragraph presents the equations for the maximum likelihood estimates of the log odds model and confidence limits when each crack is inspected only once. Maximum likelihood estimates for multiple inspections of each crack could also be developed. Further, maximum likelihood estimators for parameters of models other than equation (5) could be developed, but the solutions would not necessarily be in closed form.

Maximum likelihood estimation is based on the concept that the data will take on values which are most likely to occur under the chosen probability model. For example, in a simple Bernoulli trial (which is the probabilistic representation of a single inspection) the probability of success is  $p$ . If  $p > \frac{1}{2}$  a success would be more likely than a failure. Conversely if a success were observed in one trial, it is more likely that  $p > \frac{1}{2}$ . In following the philosophy of maximum likelihood estimation, the value of the unknown parameter that would give rise to the highest probability of obtaining the observed data is used as the estimate. In the simple Bernoulli experiment if  $p$  were equal to 1 the probability of observing a success would be 1. Since probability cannot exceed 1, the maximum likelihood estimate of  $p$  when a success is observed in a single Bernoulli trial is 1.

To find the maximum likelihood estimates of equation (5) from a sample of single inspections of  $n$  cracks, the following procedure adopted from Reference 5 can be used. The maximum likelihood estimates  $\hat{\alpha}$  and  $\hat{\beta}$  of  $\alpha$  and  $\beta$  satisfy the simultaneous equations.

$$0 = \sum_{i=1}^n z_i - \sum_{i=1}^n \frac{\exp(\hat{\alpha} + \hat{\beta} \ln(a_i))}{1 + \exp(\hat{\alpha} + \hat{\beta} \ln(a_i))} \quad (17)$$

$$0 = \sum_{i=1}^n z_i \ln(a_i) - \sum_{i=1}^n \frac{\ln(a_i) \exp(\hat{\alpha} + \hat{\beta} \ln(a_i))}{1 + \exp(\hat{\alpha} + \hat{\beta} \ln(a_i))} \quad (18)$$

where  $z_i = 1$  if the flaw is detected and 0 if it is not. The variances and covariance of the estimates  $\hat{\alpha}$  and  $\hat{\beta}$  are

$$\text{Var}(\hat{\alpha}) = \sum_{i=1}^n \frac{\exp(\alpha + \beta \ln(a_i))}{(1 + \exp(\alpha + \beta \ln(a_i)))^2} \quad (19)$$

$$\text{Var}(\hat{\beta}) = \sum_{i=1}^n \frac{(\ln(a_i))^2 \exp(\alpha + \beta \ln(a_i))}{(1 + \exp(\alpha + \beta \ln(a_i)))^2} \quad (20)$$

$$\text{Cov}(\hat{\alpha}, \hat{\beta}) = \sum_{i=1}^n \frac{\ln(a_i) \exp(\alpha + \beta \ln(a_i))}{(1 + \exp(\alpha + \beta \ln(a_i)))^2} \quad (21)$$

Estimates of these variances and covariance are calculated by substituting the estimates,  $\hat{\alpha}$  and  $\hat{\beta}$ , in equations (19), (20), and (21).

The maximum likelihood estimate of the POD function is calculated by substituting  $\hat{\alpha}$  and  $\hat{\beta}$  for  $\alpha$  and  $\beta$  in equation (5). The change of variables must be made using the same transformation that was used for the log odds model in Section 3.2 to obtain

$$\log(p(a)/(1-p(a))) = Y(a) = \hat{\alpha} + \hat{\beta} \ln(a_i) \quad (22)$$

For very large sample sizes, estimates of the variances and covariances of  $\hat{\alpha}$  and  $\hat{\beta}$  can be used to calculate a lower confidence bound on  $Y(a)$  as given by

$$Y_L(a) = \hat{\alpha} + \hat{\beta} \ln(a) - Z_\gamma \sqrt{S_{\hat{\alpha}}^2 + 2 \ln(a) S_{\hat{\alpha}\hat{\beta}}^2 + (\ln(a))^2 S_{\hat{\beta}}^2} \quad (23)$$

where

$\gamma$  is the confidence level,

$Z_\gamma$  satisfies  $P(Z < Z_\gamma) = \gamma$  for the standard normal distribution

$S_{\hat{\alpha}}^2$  is the estimate of  $\text{Var}(\hat{\alpha})$ ,

$S_{\hat{\alpha}\hat{\beta}}^2$  is the estimate of  $\text{Cov}(\hat{\alpha}, \hat{\beta})$ ,

$S_{\hat{\beta}}^2$  is the estimate of  $\text{Var}(\hat{\beta})$ .

Since the log odds transformation is monotonic, the reverse transformation of the confidence bound on  $Y(a)$  will be the confidence bound on  $P(a)$ .

An example of the use of the maximum likelihood estimates to obtain a lower bound on  $POD(a)$ , is presented in Figure 9. In this figure, the lower 95 percent confidence bound from the maximum likelihood analysis is plotted along with the optimized probability method (binomial model) that was previously presented as Figure 5. This analysis was based on the results of the inspections of 361 cracks and, with this large sample size, the 95 percent confidence limit is very closed to the mean  $POD(a)$ .

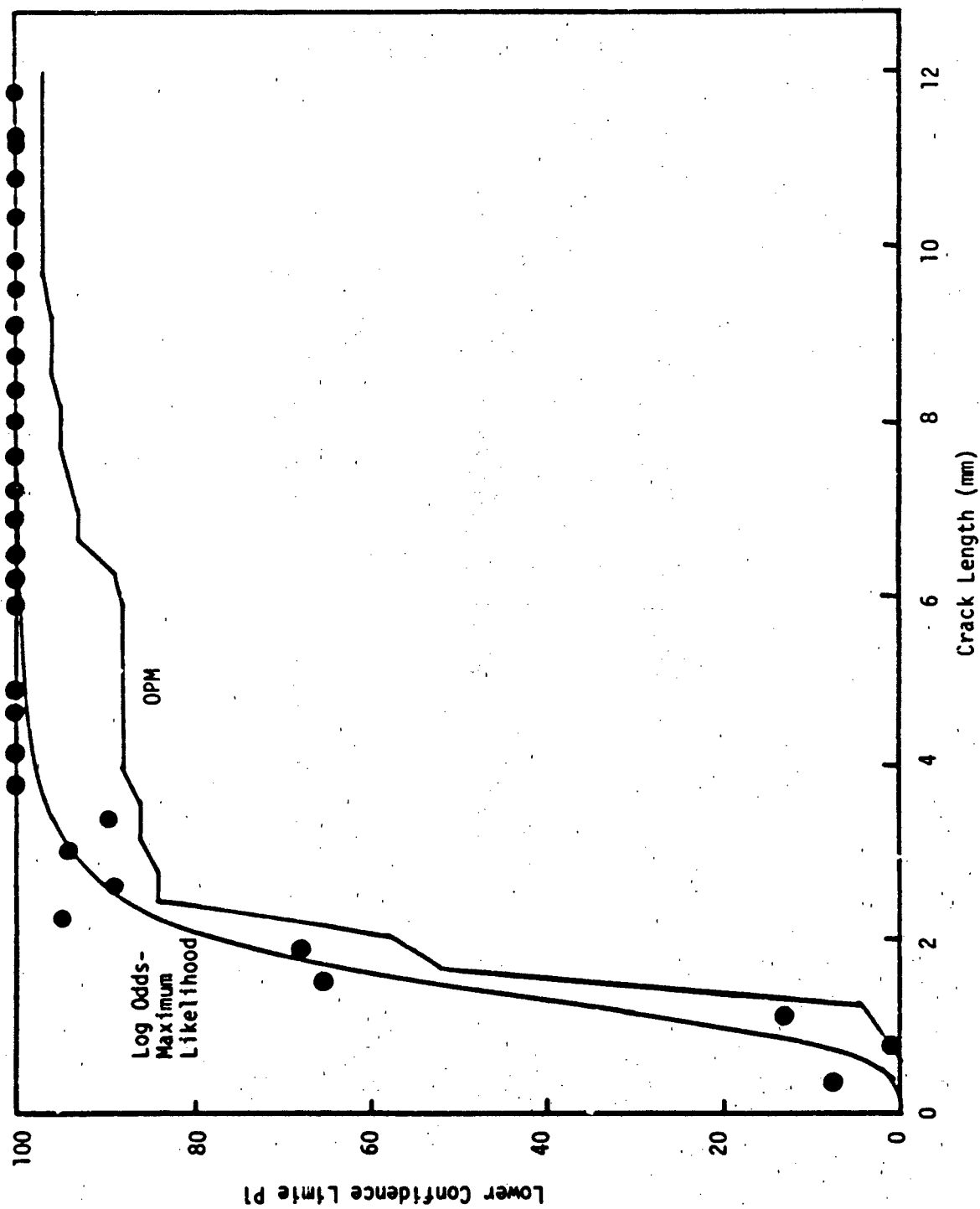


Figure 9. Example 95% Confidence Limit Using Log Odds-Maximum Likelihood Analysis.

## SECTION 4

### EVALUATION OF NDE ANALYSIS METHODS

To evaluate and compare the various methods for analyzing data from NDE capability demonstration programs, results from a large number of experiments under known conditions are necessary. Since such tests are expensive and the experimental conditions are difficult to hold fixed over the long intervals necessary to repeat experiments, "experimental" NDE data were generated in a computer simulation of inspections. The simulation process enabled the generation of a large number of NDE experiments under a "known capability" and under selected changes in experimental conditions. In the following paragraphs, the simulation process is described and the results of the evaluation of the analysis methods are presented.

#### 4.1 SIMULATED NDE CAPABILITY EXPERIMENTS

A computer program was written to simulate the NDE inspection capabilities that were present in the "Have Cracks" data of Reference 3. The following paragraphs present a general description of the program and describe the NDE experimental conditions which were simulated. The simulation program comprises Volume II of this report.

##### 4.1.1 Simulation Program

Figure 10 presents a flow diagram for the process of simulating NDE experiments. The process simulates one NDE experiment by simulating the results of one inspection of each of 400 details with cracks of different lengths. The simulation of the inspection of each detail requires three steps:

- 1) To reflect the random nature of the crack sizes that may be present in the details to be inspected, a distribution of crack sizes is assumed. A simulated inspection is initiated by selecting a crack size at random,  $a_1$ , from this distribution. The assumed crack size distribution is considered to be an experimental condition as will be discussed later.

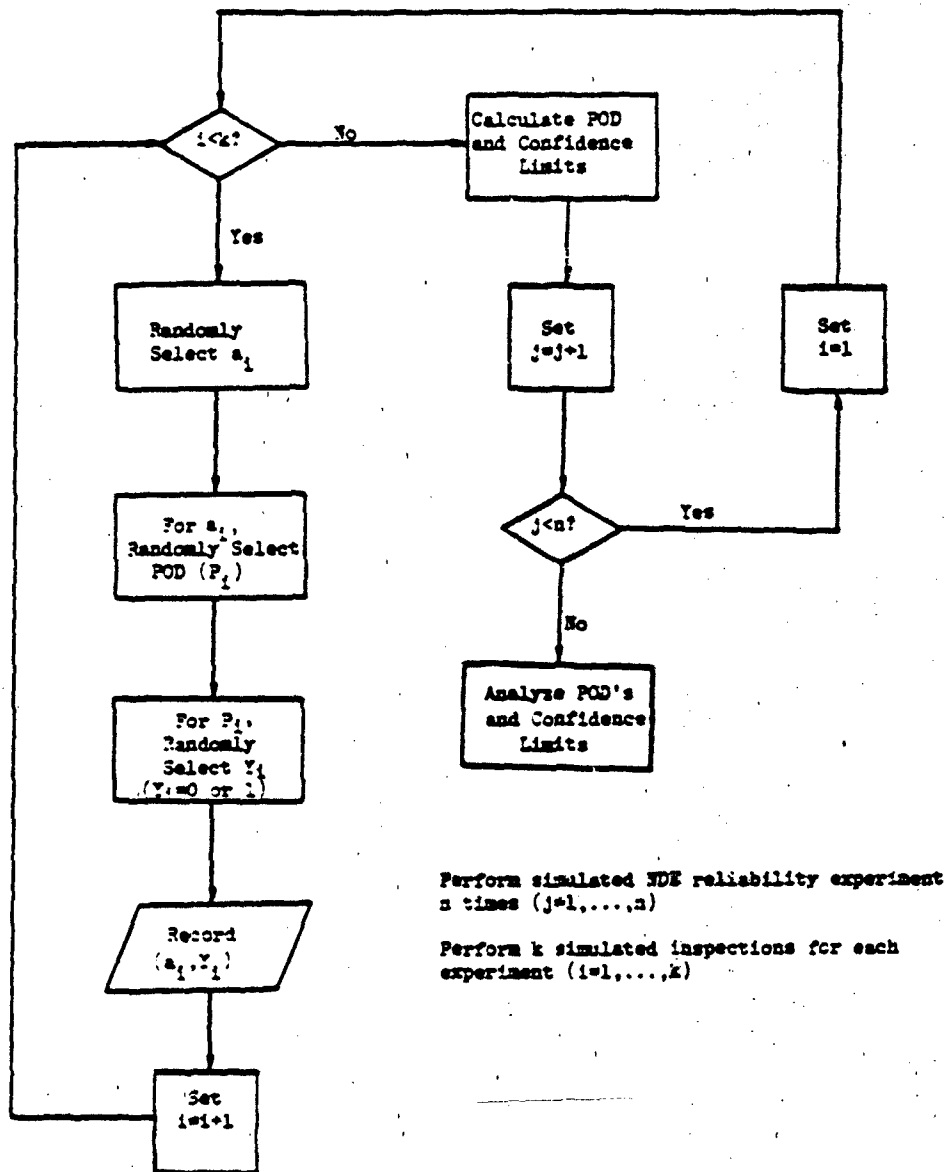


Figure 10. Flow Diagram of NDE Reliability Simulation.

2) A detection probability is randomly determined for the crack of length  $a_i$  using the regression relation between transformed crack sizes and detection probabilities for the log odds model. In particular, for the log odds model,  $a_i$  and  $p_i$  are related by the expression:

$$\ln (p_i/(1-p_i)) = \alpha + \beta \ln a_i + \epsilon_i \quad (24)$$

where  $\alpha$  and  $\beta$  are constants which specify the POD function and  $\epsilon_i$  is the difference between the average POD values at crack length  $a_i$  and the  $p_i$  value for a particular crack. It was shown that for the log odds model, the  $\epsilon_i$  can be considered to come from a normal distribution with a common standard deviation,  $S(\epsilon)$ , at all crack lengths. To randomly determine a detection probability for the crack length  $a_i$  (and given POD as defined by  $\alpha$  and  $\beta$ ),  $\epsilon_i$  is selected at random from a normal distribution with standard deviation  $S(\epsilon)$  and added to  $\alpha + \beta \ln a_i$ . The value of  $p_i$  is given by:

$$p_i = \frac{\exp(\alpha + \beta \ln a_i + \epsilon)}{1 + \exp(\alpha + \beta \ln a_i + \epsilon)} \quad (25)$$

(For the simulations of this study,  $\alpha$ ,  $\beta$ , and  $S(\epsilon)$  were taken as the values estimated for the AET, AUT, and BET data sets of the "Have Cracks Will Travel" Program.)

3) Given the detection probability,  $p_i$ , for the crack, a simple Bernoulli trial is simulated with probability  $p_i$  of successfully detecting the crack and  $(1-p_i)$  of failing to detect it. The result of the "inspection" is recorded either as  $(a_i, 1)$  if the "crack" was "detected" or  $(a_i, 0)$  if the "crack" was not "detected."

After the above steps are repeated 400 times to complete an entire experiment, the data were analyzed by seven analysis procedures. These included the four interval methods based on the binomial distribution (60 points per interval and 50 percent overlap), a slightly modified version of the Lockheed analysis of Reference 3, and the regression and maximum likelihood analyses using the log odds model. Lower confidence limits on probability



of detection (upper POD/CL limits) were calculated using all combinations of POD equal to 0.5, 0.75, 0.9, 0.95, and 0.99 and confidence limits for 90, 95, and 99 percent. In selected runs a 50 percent confidence limit was also calculated as a method of generating the mean POD curve. These estimates were recorded for later analysis while the results of the individual "inspections" were discarded.

The above procedure was repeated as part of the simulation process to generate a large number of repetitions of the basic NDE reliability experiment and the associated estimates of the POD. These data form the basis for the comparison and evaluation of the various methods of estimating the capability of an NDE system.

Volume II contains a complete description of the simulation program, instructions for the use of the program and a complete source listing. Input parameters and output format are also described. The output comprises a file of the POD/CL limits for all samples generated in the run and tables which summarize the performance of the various POD/CL limits. Further analysis of the POD/CL limits was performed using a standard statistical program package.

#### 4.1.2 Experimental Conditions and Results

Table 3 summarizes the experimental conditions which were simulated. Each condition is identified as an experimental "environment" which is defined in terms of a POD capability, a crack size distribution and a standard error,  $S(e)$ . The  $\alpha$  and  $\beta$  values of POD curves were those which resulted from the analysis of the indicated "Have Cracks" data sets. These capabilities are displayed graphically in Figure 11. Crack length distributions B and A were the empirical distributions observed in the sample B and sample A specimens of the "Have Cracks" study.

In order to generate data that was more compatible with the interval analysis methods, it was necessary to introduce a distribution of longer cracks. This distribution is identified as L and was defined as a lognormal distribution with a median of 12.7mm and the 90th percentile at 76.2mm. The long cracks were used only with the AET capability.

TABLE 3  
DEFINITION OF EXPERIMENTAL CONDITIONS  
USED IN SIMULATED NDE EXPERIMENTS

Environment Code	Crack Length Distribution	POD CURVE			
		Parameters		Percentiles	
		$\alpha$	$\beta$	a0.9	a0.95
BET (1)	B	-0.65	0.88	25.4	59.4
AUT (2)	A	-3.9	1.8	29.6	44.8
AET (3)	A	-2.9	1.7	20.1	31.1
AET-L.C.	L	-2.9	1.7	20.1	31.1
AET-L.C. - Low Scatter	L	-2.9	1.7	20.1	31.1

- (1) Representative of eddy current surface scans around countersunk fasteners, skin and stringer wing segments (Sample B).
- (2) Representative of ultrasonic shear wave scans around countersunk fasteners, skin and stringer wing assembly (Sample A).
- (3) Representative of eddy current surface scans around countersunk fasteners, skin and stringer wing assembly (Sample A).

A reduced value of the standard error of deviations from the transformed model was introduced to demonstrate the effect of a stronger correlation of detection probability with crack length on the estimates of POD/CL. This comparison was made only for the AET long cracks environment. In particular, it was assumed for the AET-L.C.-low scatter environment that  $S(e)$  was one-fourth of the value observed in the AET "Have Cracks" data.

#### 4.2 EVALUATION OF RESULTS

The initial objectives for the evaluation of the data from the simulated experiments were to compare the various methods of analyzing NDE reliability data and to determine the effect of the choice of POD/CL combination on the reliability characterization. These evaluations are in the following paragraphs. However, during the study, it was observed that the length of cracks in the NDE experiment can influence the characterization and that the

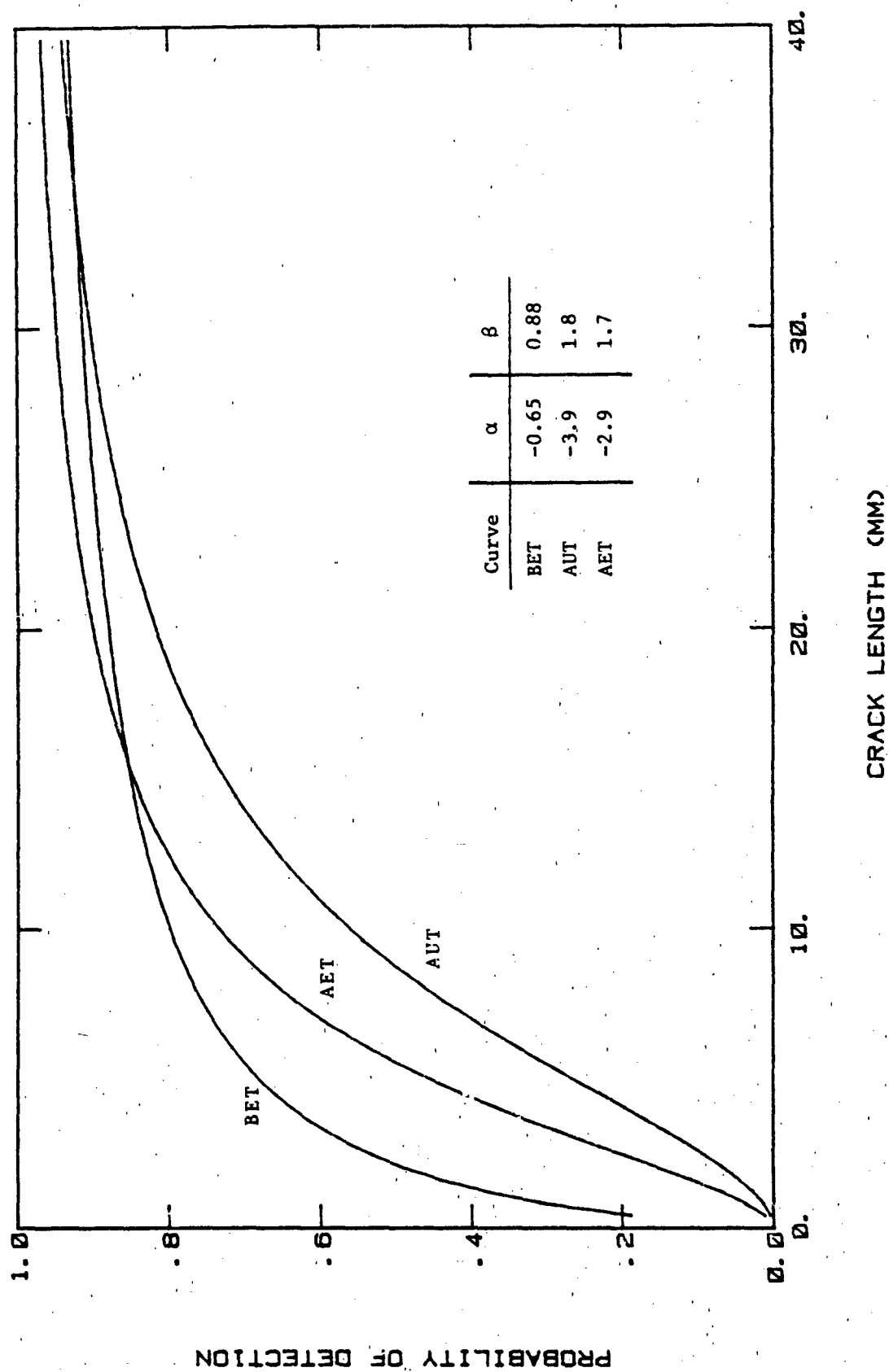


Figure 11. POD Curves Used in the Simulation Study.

strength of the correlation of detection probability with crack length does not have as much effect as would be expected. These latter evaluations are also presented in the following.

#### 4.2.1 Comparison of Analysis Methods

In a preliminary evaluation of the seven analysis methods, 25 experiments were simulated under the AET environment which is based on the distribution of crack lengths that were present in the "Have Cracks" data. Table 4 presents percentages of three categories of events derived from the 25 simulations:

1) the percent of times the confidence bound was greater than the true values; 2) the percent of time the confidence bound crossed the POD more than once; and 3) the percent of times the confidence bound never reached the POD value. For this crack length distribution, the interval methods always failed to yield estimates of the 90/95 and 95/90 crack length. That is, for the range of crack lengths in the experiment and the capability determined for the AET POD function, it was extremely unlikely to obtain any interval which would yield sufficient detections to have 95 percent confidence that 90 percent of all cracks of that length would be found.

To effect a comparison between the interval and regression model approaches to the analysis, a distribution of longer cracks was introduced into the simulation. In particular, it was assumed that the crack lengths that would be present in the NDE experiment were from a lognormal distribution with a median crack length of 12.7mm and a 90th percentile of 76.2mm. The particular choice of lognormal family was judged to be inconsequential and the percentiles were selected to yield much longer cracks than those of the original "Have Cracks" Sample A data. All cracks in the Sample A data were less than 26mm and approximately 90 percent were less than the 12.7mm assumed for the median of the lognormal crack size distribution. Table 5 presents percentages analogous to those of Table 4 but with the longer cracks size distribution and for 75 simulated experiments.

TABLE 4  
PRELIMINARY EVALUATION OF VARIOUS ESTIMATION PROCEDURES  
BASED ON 25-AET SIMULATED EXPERIMENTS

METHOD	BOUND HIGHER THAN TRUE %		DOUBLE VALUES %		NO ESTIMATE %	
	90/95	95/90	90/95	95/90	90/95	95/90
RIM	100*	100*	0	0	100	100
NOSPM	100*	100*	0	0	100	100
OSPM	100*	100*	0	0	100	100
OPM	100*	100*	0	0	100	100
LOCKHEED	100	100	0	0	0	0
LOG ODDS -REGRESSION	92	84	0	0	0	0
LOG ODDS -MAX. LIKE.	84	84	0	0	0	0
IDEAL	95	90	0	0	0	0

\*When an estimate could not be calculated the POD/CL limit was considered to be infinity and therefore higher than the true value so it was included in this category.

TABLE 5  
EVALUATION OF VARIOUS ESTIMATION PROCEDURES BASED  
ON 75 SIMULATED EXPERIMENTS - AET- LONG CRACKS

	BOUND HIGHER THAN TRUE %		DOUBLE VALUES %		NO ESTIMATE %	
	90/95	95/90	90/95	95/90	90/95	95/90
RIM	100*	100*	17.3	0	82.7	100
NOSPM	100*	100*	5.3	0	41.3	76
OSPM	100*	100*	9.3	0	37.3	76
OPM	100*	100*	4	5.3	13.3	73.3
LOCKHEED	100	100	0	0	0	0
LOG ODDS -REGRESSION	100	100	0	0	0	0
LOG ODDS -MAX. LKHD.	85.3	77.3	0	0	0	0
IDEAL	95	90	0	0	0	0

\* When an estimate could not be calculated the POD/CL limit was considered to be infinity and therefore higher than the true value so it was included in this category.

Even with the longer crack size distributions, the binomial analyses of the interval methods often do not provide a limit. Note that the true 90 and 95 percent POD values for this experiment are 20.1 and 31.1mm, respectively, as these were pre-determined by the manner in which the simulation was performed. Double values occurred in the interval estimates, even for the Optimized Probability Method (OPM) which is designed to produce a better behaved POD function.

All of the methods except the log odds-maximum likelihood produced POD/CL estimates which were always greater than the true POD value. The row labeled ideal indicates the expected percentages higher than true. While the calculated values are conservative, the fact that all exceed the true value indicates that they are not true confidence limits on the POD value. The log odds-maximum likelihood analysis on the other hand produced too many samples which were below the true value. The cause of these lower than desired values is unknown but it is postulated that this result is due to the transformation in the simulation procedure generating the correct median POD function rather than the correct mean POD function.

A more detailed method for comparing the POD/CL estimates from the different analysis methods is to compare the observed distributions of the estimates that were obtained during the simulated experiments. Figures 12 and 13 present the observed cumulative distributions of the 90/95 and 95/90 values, respectively, for the OPM, log odds-regression, and log odds-maximum likelihood analysis methods. These figures are based on the AET-long cracks environment and clearly demonstrate that the OPM estimates of the 90/95 and 95/90 are generally much greater than those of the log odds regression estimates which are in turn much greater than those of the log odds-maximum likelihood estimates. The failure of the OPM distributions to reach unity reflects the previously noted percentage of experiments for which no estimate of the POD/CL combination is reached using the binomial distribution.

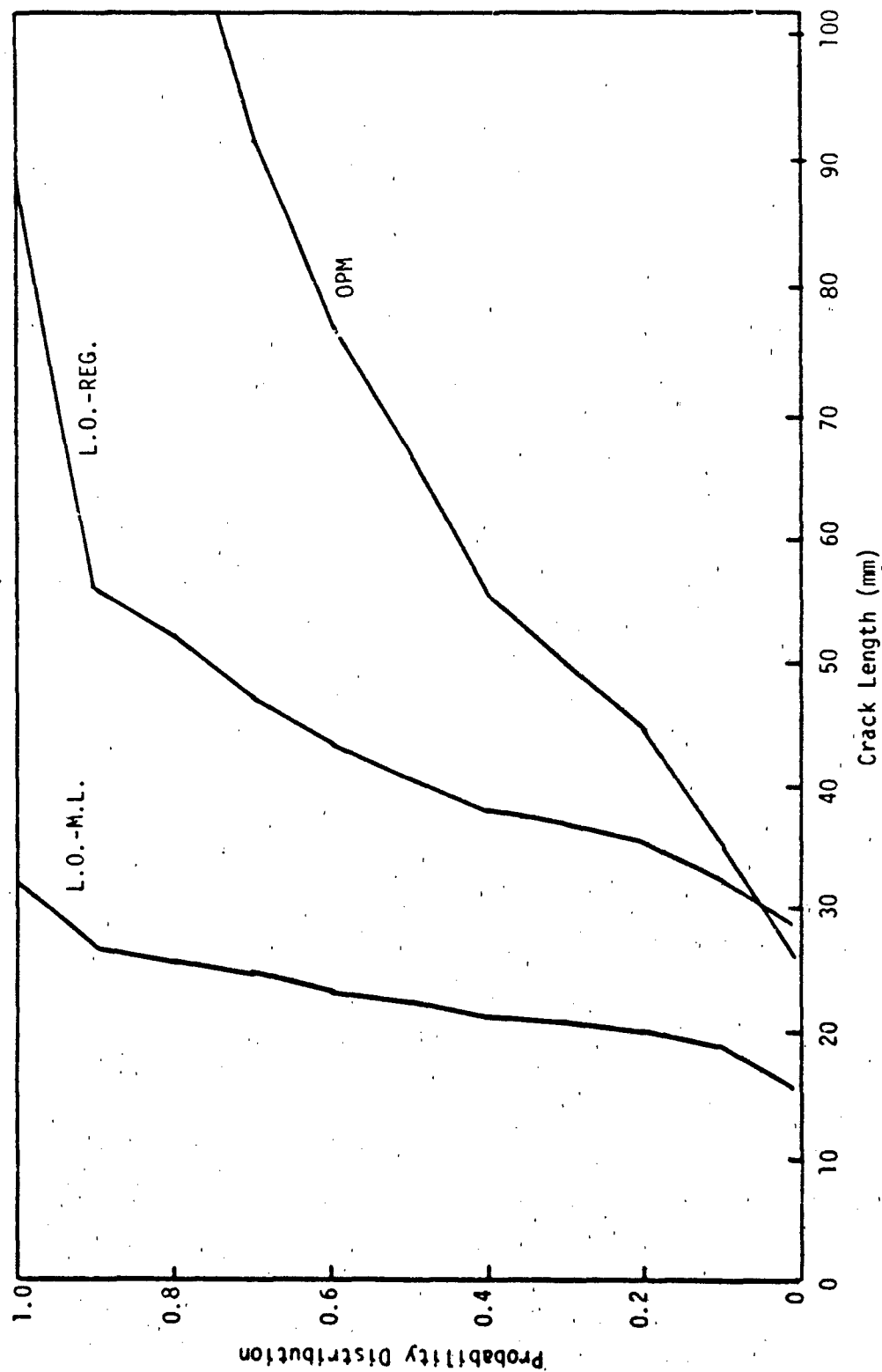


Figure 12. Comparison of Distributions of 90/95 Limits Produced by the Log Odds-Maximum Likelihood, Log Odds-Regression, and Optimized Probability Methods in Environment AET - Long Cracks.



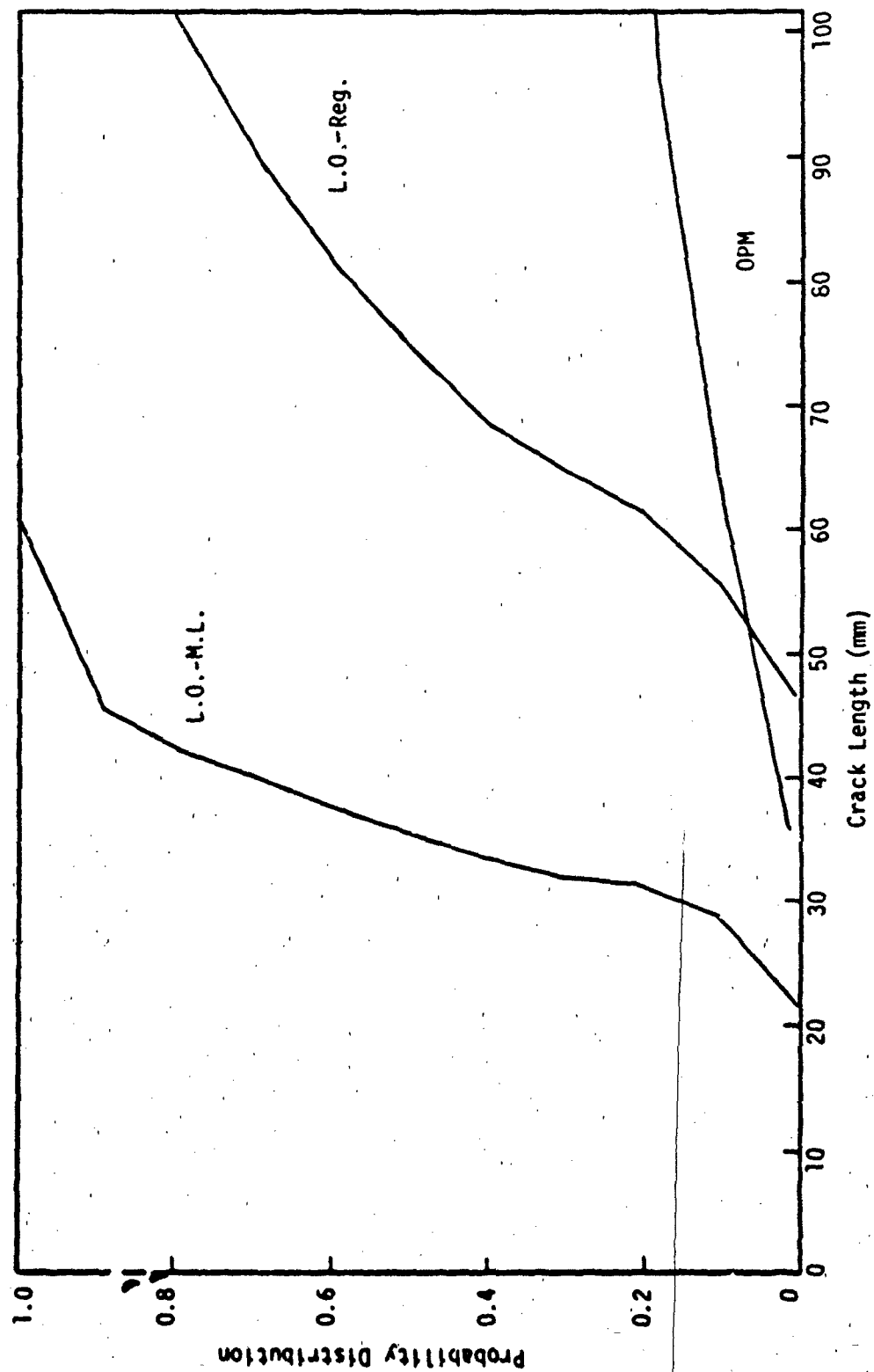


Figure 13. Comparison of Distributions of 95/90 Limits Produced by the Log Odds-Maximum Likelihood Log Odds-Regression, and Optimized Probability Methods in Environment AEI-Long Cracks.

The slope of the distribution functions reflect the scatter that can be expected in future estimates using the various analysis methods. The smaller the slope, the greater is the scatter and lack of precision in the estimates. For example, referring to the OPM curve of Figure 12, the likely range of a future 90/95 limit is quite large. For example, there is a 50 percent chance that a future estimate would be either less than 47mm or greater than 100mm (a factor of 2). From the log odds-maximum likelihood curve, however, the corresponding middle of 50 percent spans the range of 21mm to 25mm, a range of a factor of 1.2. The precision of log odds-regression analyses is between these two.

The regression approach to estimating confidence limits for high level POD values yields limits that are both closer to the true limit and provide a more precise (in the sense of less scatter) estimate. This is clearly demonstrated for the OPM method which has been shown to be the "best" of the interval methods.

Figures 14 through 19 compare the distributions of 90/95 and 95/90 limits for the log odds-maximum likelihood, log odds-regression, and Lockheed analysis methods for the BET, AUT, and AET environments. These environments have the shorter crack size distributions that were present in the Sample B and Sample A "Have Cracks" data. All of these figures display that the Lockheed analysis yields significantly larger estimates of the 90/95 and 95/90 estimates and that these estimates have less precision than those of the log odds analysis model. This result was expected since the underlying simulation model was the log odds equation and since the Lockheed analyses placed confidence bounds on the individual detection probabilities at a crack length rather than on the mean of the detection probabilities.

The log odds-maximum likelihood estimates are smaller and have greater precision than those of the log odds-regression for all six combinations of Figures 14 through 19. For these shorter crack simulations, the confidence limits are closer to true in the sense that they are less than the true value in the correct

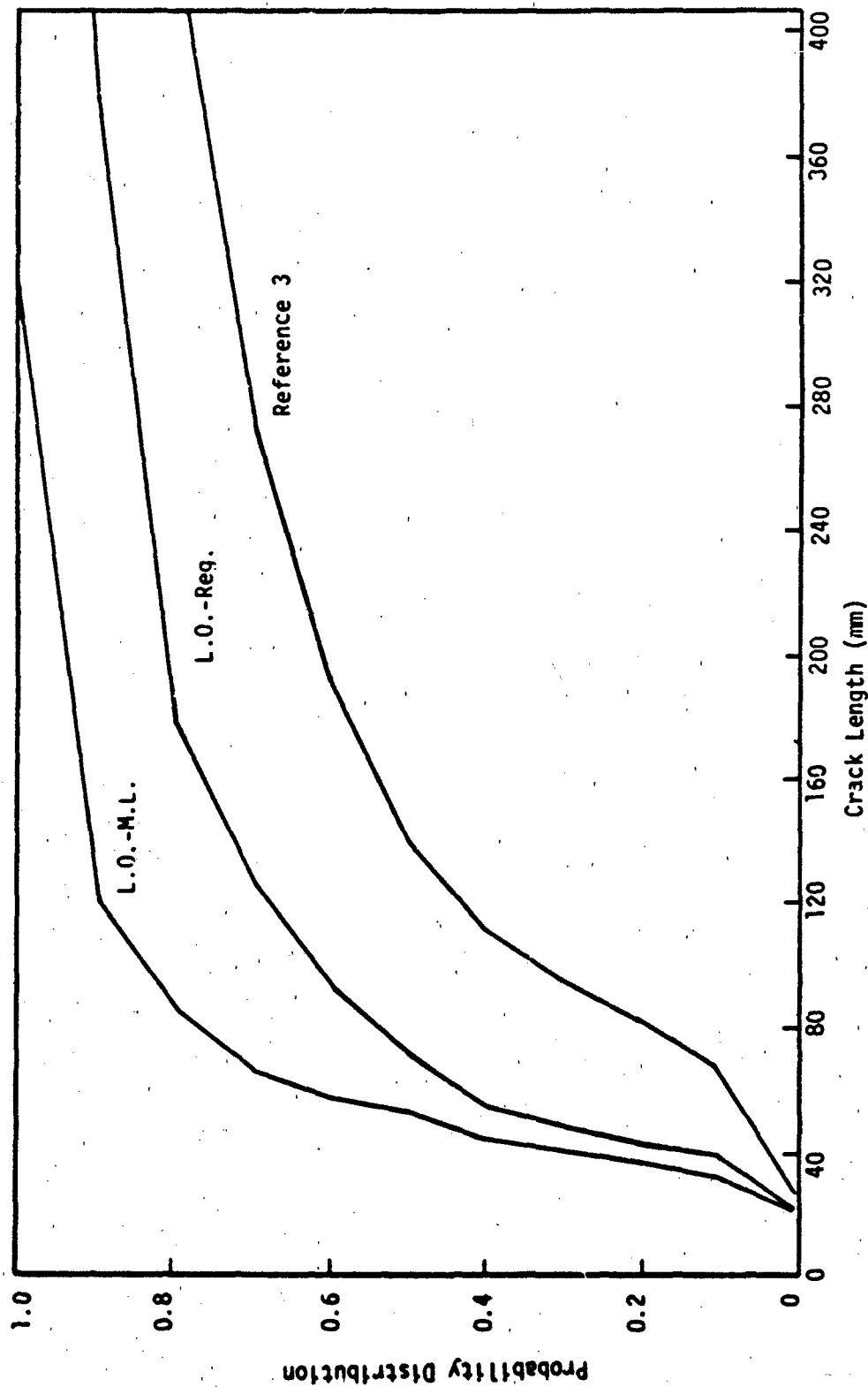


Figure 14. Comparison of Distributions of 90/95 Limits Produced by the Log Odds-Maximum Likelihood, Log Odds-Regression and Reference 3 Methods in Environment BET.

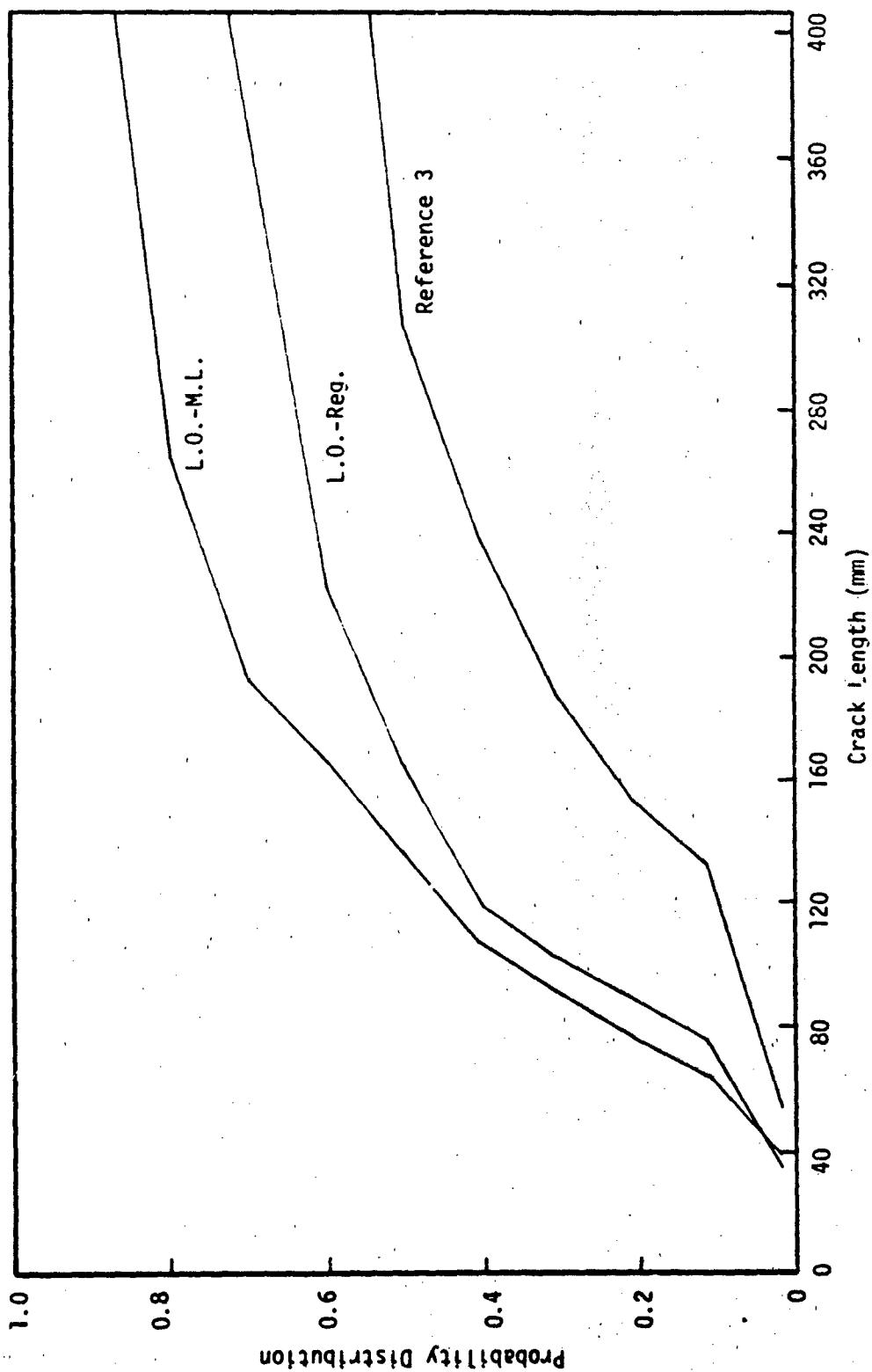


Figure 15. Comparison of Distributions of 95/90 Limits Produced by the Log Odds-Maximum Likelihood, Log Odds-Regression, and Reference 3 Methods in Environment BET.

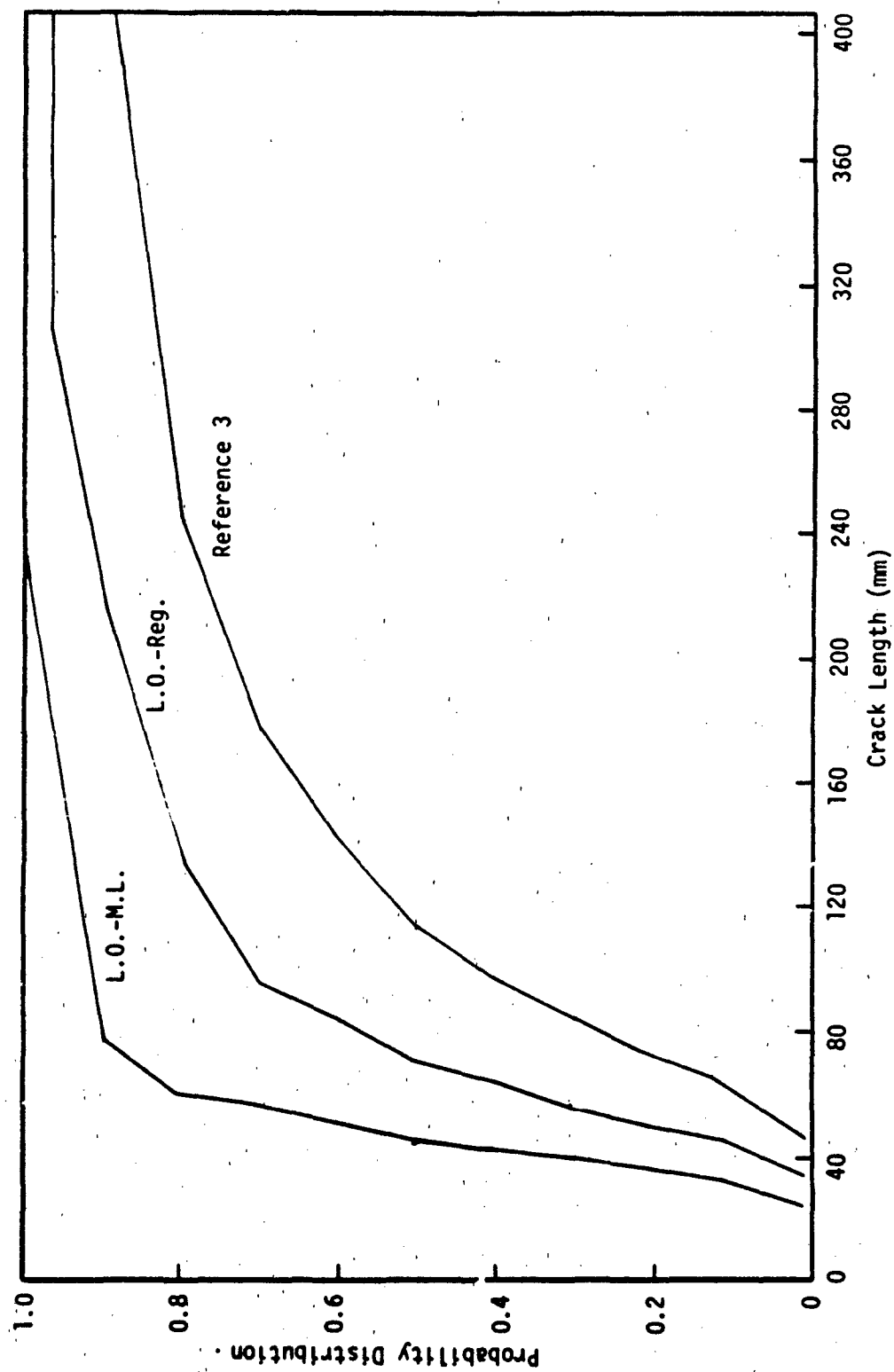


Figure 16. Comparison of Distributions 90/95 Limits Produced by the Lod Odds-Maximum Likelihood, Log Odds-Regression, and Reference 3 Methods in Environment AUT.

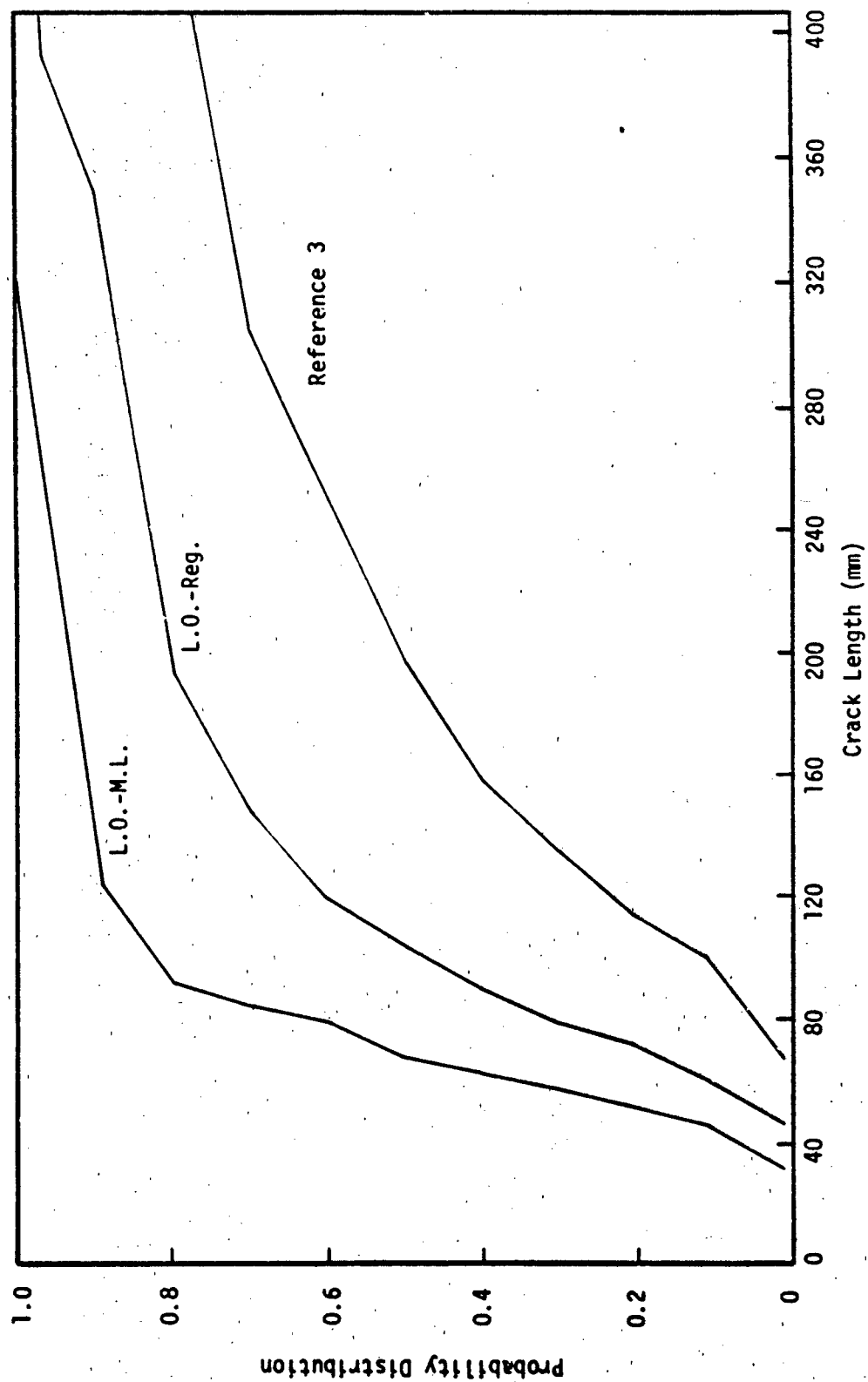


Figure 17. Comparison of Distributions of 95/90 Limits Produced by the Log Odds-Maximum Likelihood, Log Odds-Regression, and Reference 3 Methods in Environment AUT.

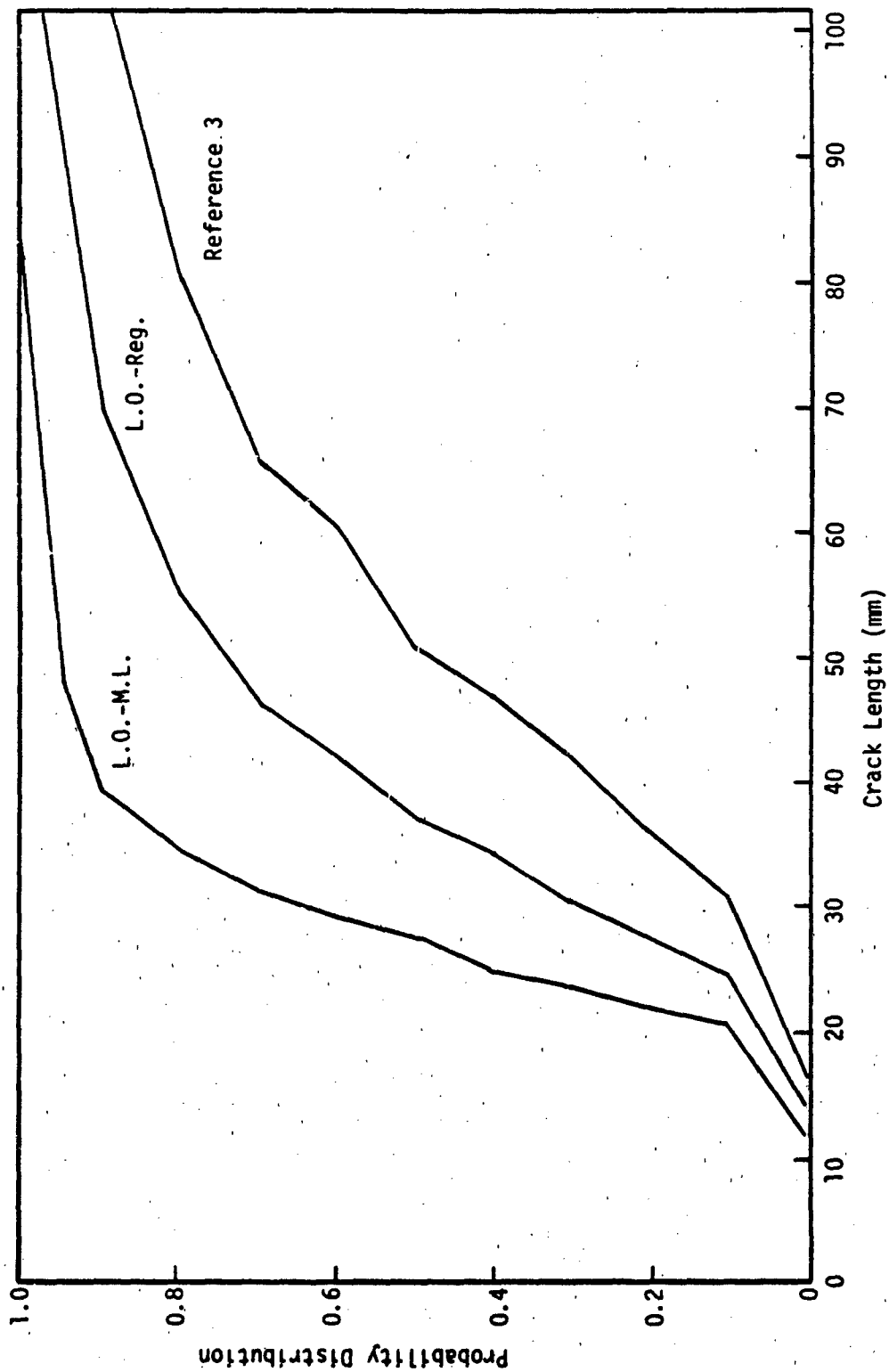


Figure 18. Comparison of Distributions of 90/95 Limits Produced by the Log Odds-Maximum Likelihood, Log Odds-Regression and Reference 3 Methods in Environment  $\text{AET}$ .

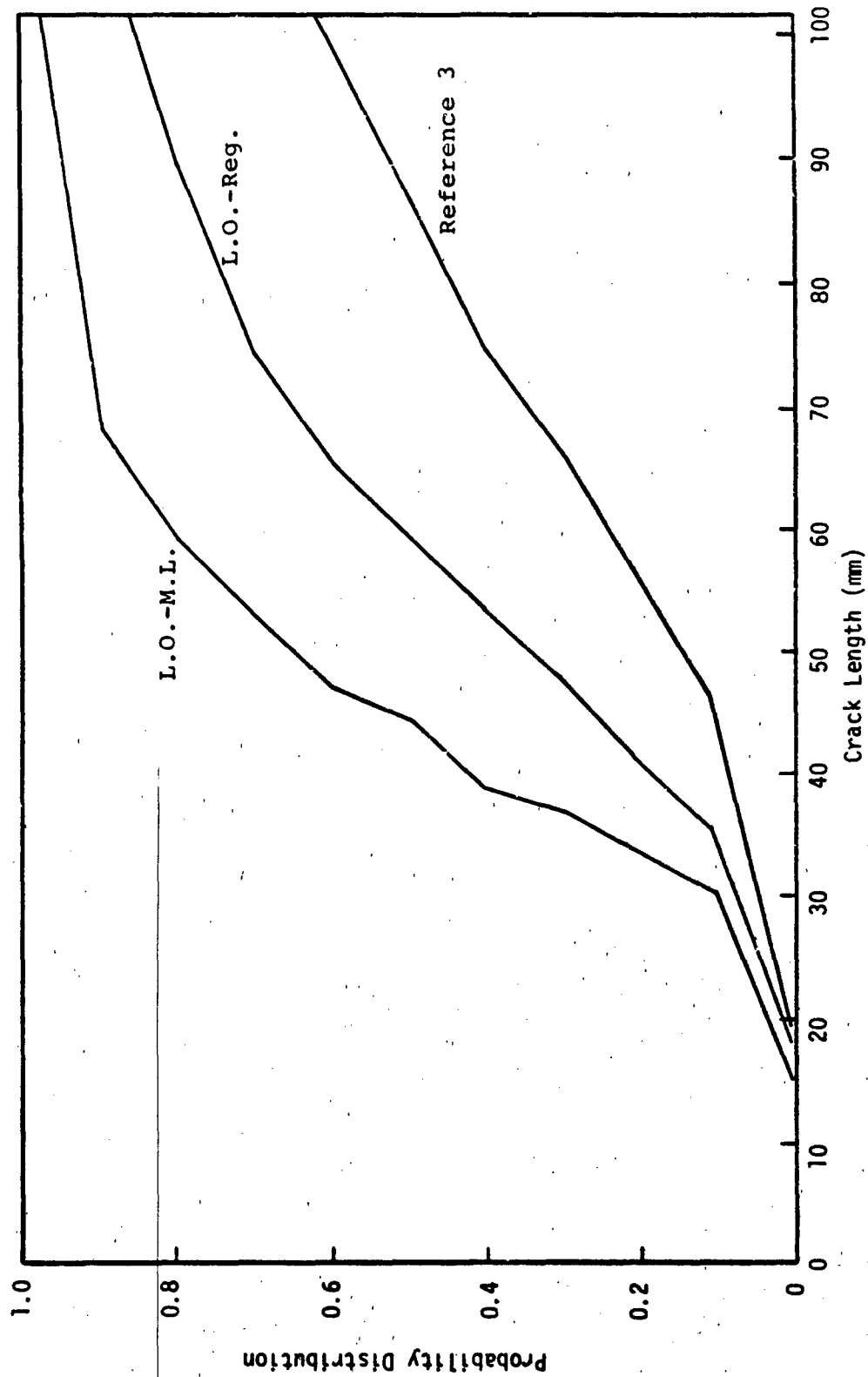


Figure 19. Comparison of Distributions of 95/90 Limits for Log Odds-Maximum Likelihood, Log Odds-Regression and Reference 3 Methods in Environment AET.



proportion of experiments. However, the precision of the estimates for any of these analysis methods is poor and will be discussed later.

#### 4.2.2 Comparison of POD/CL

The choice of the POD/CL combination to be used in defining the capability of an NDE system has been rather arbitrarily defined as 90/95. To evaluate various choices from the viewpoint of their estimates in an NDE evaluation experiment, the crack lengths corresponding to several combinations of POD and confidence level were calculated for each simulated experiment. The statistical properties of these POD/CL limits under fixed conditions provided considerable insight into the practical usefulness of various combinations.

Tables 6, 7, and 8 present the mean ( $\bar{X}$ ), standard deviation (S), and coefficient of variation ( $CV = 100 S/\bar{X}$ ) of the estimates of POD/CL limits for the BET, AUT, and AET environments obtained using the log odds model for analysis. The statistics are based on a sample of 100 estimates ("experiments") in each of the environments. The "true" crack length corresponding to a POD value (Equation 5 or Figure 11) is listed as the  $a_p$  value.

The coefficient of variation columns of the tables display that estimation precision decreases rapidly with increasing POD and with increasing level of confidence. Further, the average of the calculated POD/CL limits increases as the degree of confidence increases as would certainly be expected. The combination of these facts indicates that considerable real scatter is present in the estimate of NDE capability at high values of POD and confidence.

As an example of the effect of scatter in the estimates, assume an  $a_{NDE}$  value is to be determined for the BET environment and the data will be analyzed using the log odds-regression approach. Figure 20 displays the distribution of potential estimates of  $a_{NDE}$  if  $a_{NDE}$  is defined as either the 90/95 or 95/90 limit. The result of the future experiment in the BET environment is equivalent to drawing a number at random from

TABLE 6  
MEANS, STANDARD DEVIATIONS, AND COEFFICIENTS OF VARIATION FOR POD/CL  
LIMITS IN ENVIRONMENT BET ( $\alpha = -0.65$   $\beta = 0.88$ )

POD	CL	$a_p$ (mm)	LOG ODDS-REG.			LOG ODDS-ML		
			$\bar{X}$ (mm)	S (mm)	CV (%)	$\bar{X}$ (mm)	S (mm)	CV (%)
0.5	0.5	2.1	2.0	0.6	30	2.0	0.5	25
	0.9		2.8	0.6	21	2.5	0.5	20
	0.95		3.0	0.6	20	2.5	0.5	20
	0.99		*	*	704	2.8	0.5	18
0.75	0.5	7.3	8.1	1.3	16	8.4	1.3	15
	0.9		10.7	2.6	24	9.4	1.8	19
	0.95		12.2	5.3	43	9.9	1.9	19
	0.99		*	*	1000	10.7	2.4	22
0.9	0.5	25.4	34.8	17.6	51	38.4	18.0	47
	0.9		145.5	391.0	269	57.4	48.5	84
	0.95		*	*	790	68.1	75.4	111
	0.99		*	*	990	111.5	260.8	234
0.95	0.5	59.4	102.9	94.2	92	114.8	98.1	85
	0.9		*	*	541	248.4	505.0	203
	0.95		*	*	975	362.7	*	294
	0.99		*	*	990	*	*	629
0.99	0.5	387.8	*	*	243	*	*	255
	0.9		*	*	824	*	*	656
	0.95		*	*	*	*	*	813
	0.99		*	*	990	*	*	972

\* Indicates value was larger than 1,000.

TABLE 7

MEANS, STANDARD DEVIATIONS, AND COEFFICIENTS OF VARIATION FOR POD/CL  
LIMITS IN ENVIRONMENT AUT ( $\alpha = -3.9, \beta = 1.8$ )

POD	CL	$a_p$ (mm)	LOG ODDS-REG.			LOG ODDS-ML		
			$\bar{X}$ (mm)	S (mm)	CV (%)	$\bar{X}$ (mm)	S (mm)	CV (%)
0.5	0.5	8.7	9.1	0.8	9	8.9	0.7	8
	0.9		10.9	41.9	17	9.4	0.9	10
	0.95		11.9	2.8	24	9.4	1.0	11
	0.99		17.3	17.4	101	10.2	1.2	12
0.75	0.5	16.1	19.0	4.1	22	17.5	3.3	19
	0.9		29.0	16.5	57	20.6	5.4	26
	0.95		37.1	36.7	99	21.6	6.4	30
	0.99		519.7		554	24.4	9.4	39
0.9	0.5	29.6	40.4	16.6	41	35.3	11.7	33
	0.9		87.6	124.6	142	46.7	23.4	50
	0.95		161.5	436.6	270	51.6	30.1	58
	0.99		*	*	660	65.3	53.4	82
0.95	0.5	44.8	68.3	40.0	59	57.7	25.8	45
	0.9		208.8	489.2	234	83.8	59.4	71
	0.95		562.6	*	422	96.3	80.8	84
	0.99		*	*	681	134.9	166.0	123
0.99	0.5	112.1	230.1	264.7	115	174.8	133.9	77
	0.9		*	*	496	328.4	437.6	133
	0.95		*	*	671	422.1	688.6	163
	0.99		*	*	709	798.8	*	255

\* Indicates value was larger than 1,000.

TABLE 8  
MEAN, STANDARD DEVIATIONS, AND COEFFICIENTS OF VARIATION FOR POD/CL  
LIMITS IN ENVIRONMENT AET ( $\alpha = -2.9$ ,  $\beta = 1.7$ )

POD	CL	$a_p$ (mm)	LOG ODDS-REG.			LOG ODDS-ML		
			$\bar{X}$ (mm)	S (mm)	CV (%)	$\bar{X}$ (mm)	S (mm)	CV (%)
0.5	0.5	5.5	5.6	0.4	7	5.6	0.4	7
	0.9		6.1	0.5	8	5.8	0.4	7
	0.95		6.4	0.5	8	5.8	0.4	7
	0.99		6.9	0.7	10	6.1	0.4	7
0.75	0.5	10.5	11.4	1.6	14	11.2	1.4	13
	0.9		14.2	3.0	21	12.2	2.0	16
	0.95		15.5	4.1	26	12.7	2.1	17
	0.99		21.1	12.4	59	13.5	2.5	19
0.9	0.5	20.1	24.1	6.5	27	23.1	5.6	24
	0.9		36.6	17.8	49	27.7	8.9	32
	0.95		45.0	31.8	71	29.5	10.4	35
	0.99		122.4	433.8	354	33.5	14.7	44
0.95	0.5	31.1	40.4	15.5	38	38.1	12.7	33
	0.9		72.6	58.2	80	49.3	22.6	46
	0.95		99.8	131.3	132	53.8	27.7	51
	0.99		747.3	*	691	65.3	42.9	66
0.99	0.5	82.2	130.8	90.9	69	118.1	68.3	58
	0.9		383.3	838.8	216	184.4	158.8	36
	0.95		802.9	*	409	216.4	214.9	99
	0.99		*	*	969	312.2	425.2	136

\* indicates value was larger than 1,000.

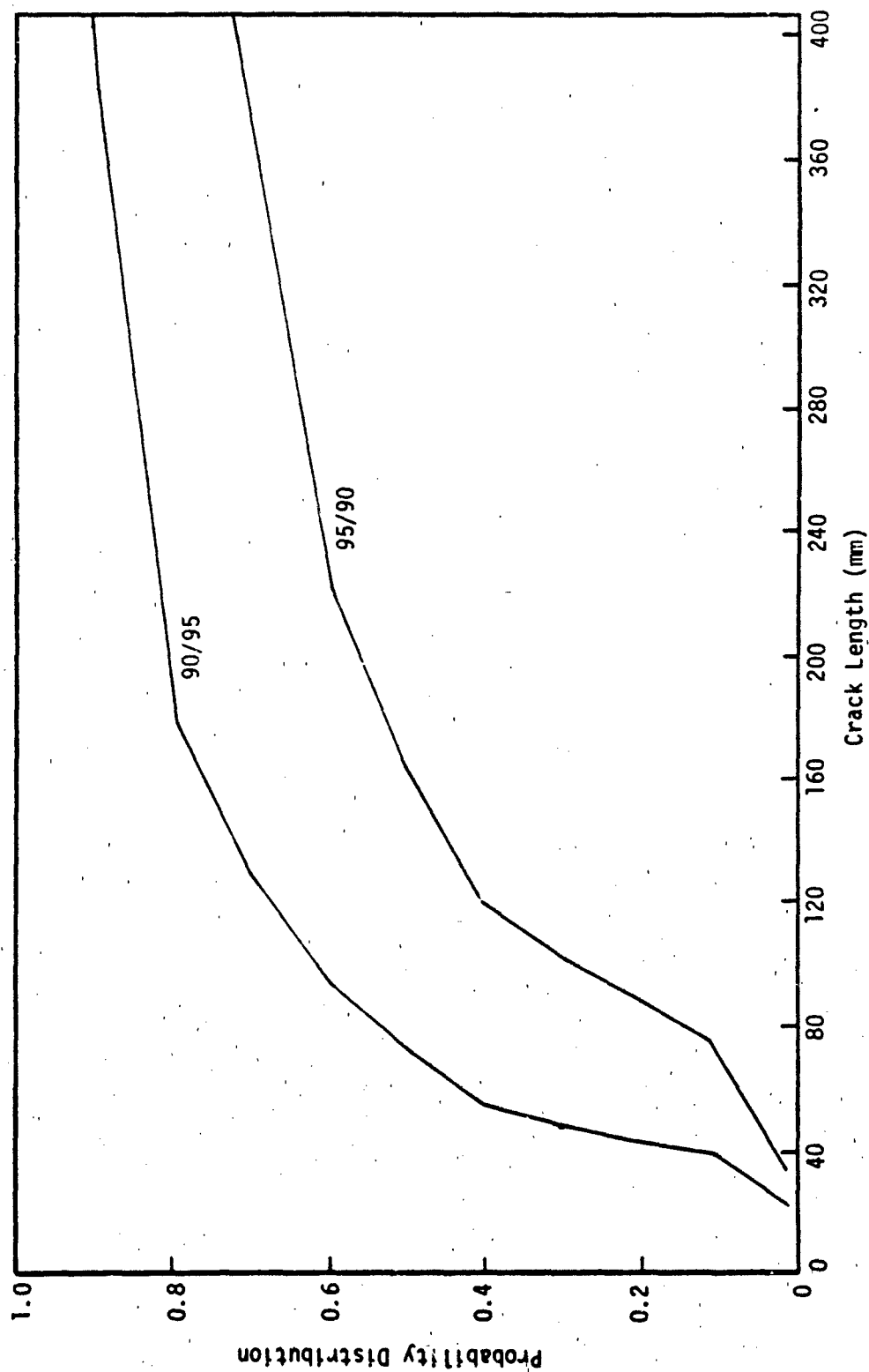


Figure 20. Comparison of Distributions of 90/95 and 95/90 Limits Produced by the Log Odds-Regression Method in Environment BET.

either of these cumulative distribution functions. For this environment, the "true" 90th percentile is 25.4mm (Table 6) but there is a 50 percent chance that the estimate will exceed 71mm, a 25 percent chance that the estimate will exceed 150mm and a 10 percent chance that the estimate will exceed 380mm. These values were obtained from the 90/95 curve of Figure 20. Similarly, the "true" 95th percentile is 59.4mm while there is a 50 percent chance that the 95/90 estimate will exceed 160mm and a 28 percent chance that this estimate will exceed 400mm. Similar examples could be presented for the other environments and the log odds-maximum likelihood analysis.

In general, the scatter in the estimates is sufficiently large as to cast considerable doubt on the validity of any single POD/CL limit if the POD is 0.9 or greater and the level of confidence is 0.9 or greater. It should be noted that the scatter in the limits gives rise to excessively large estimates of  $a_{NDE}$  and the estimates are conservative. However, the degree of conservativeness would be unknown in a particular application.

#### 4.2.3 Comparison of Crack Size Distributions in the Specimens of an NDE Capability Experiment

To compare the interval and regression model methods of analysis, it was necessary to introduce a distribution of long cracks for the representative "specimens" of the simulation. When the resulting POD/CL estimates from the long cracks simulation were compared to those of the short cracks (i.e. the crack sizes of the "Have Cracks" data) simulations, it was observed that significantly different distributions resulted. Examples of such comparisons for the AET environment are presented in Figures 21 and 22 for the log odds-regression and log odds-maximum likelihood analysis methods, respectively.

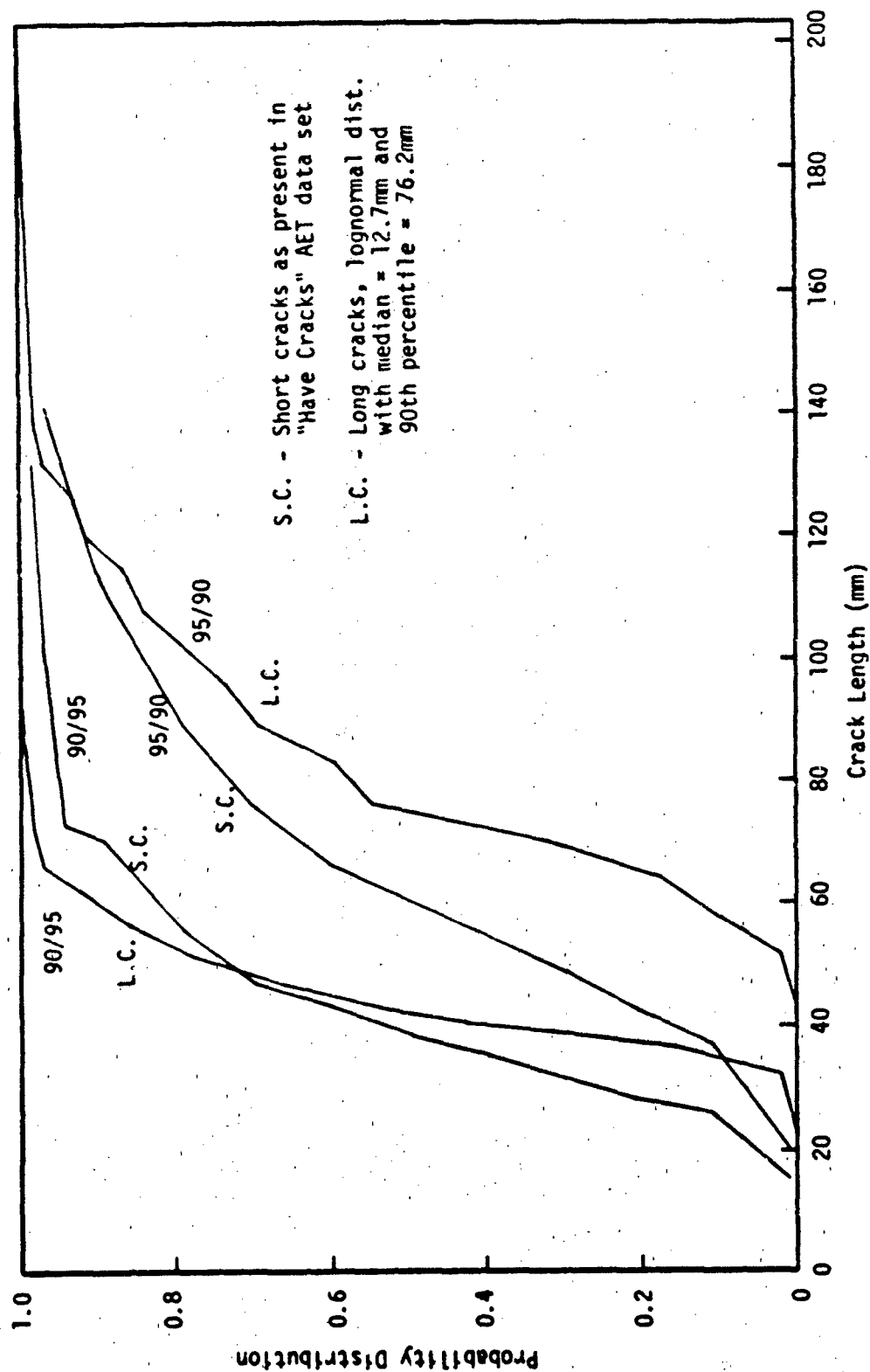


Figure 21. Comparison of Specimen Crack Sizes in an NDE Capability Experiment - AET Environment, Log Odds-Regression Analysis.

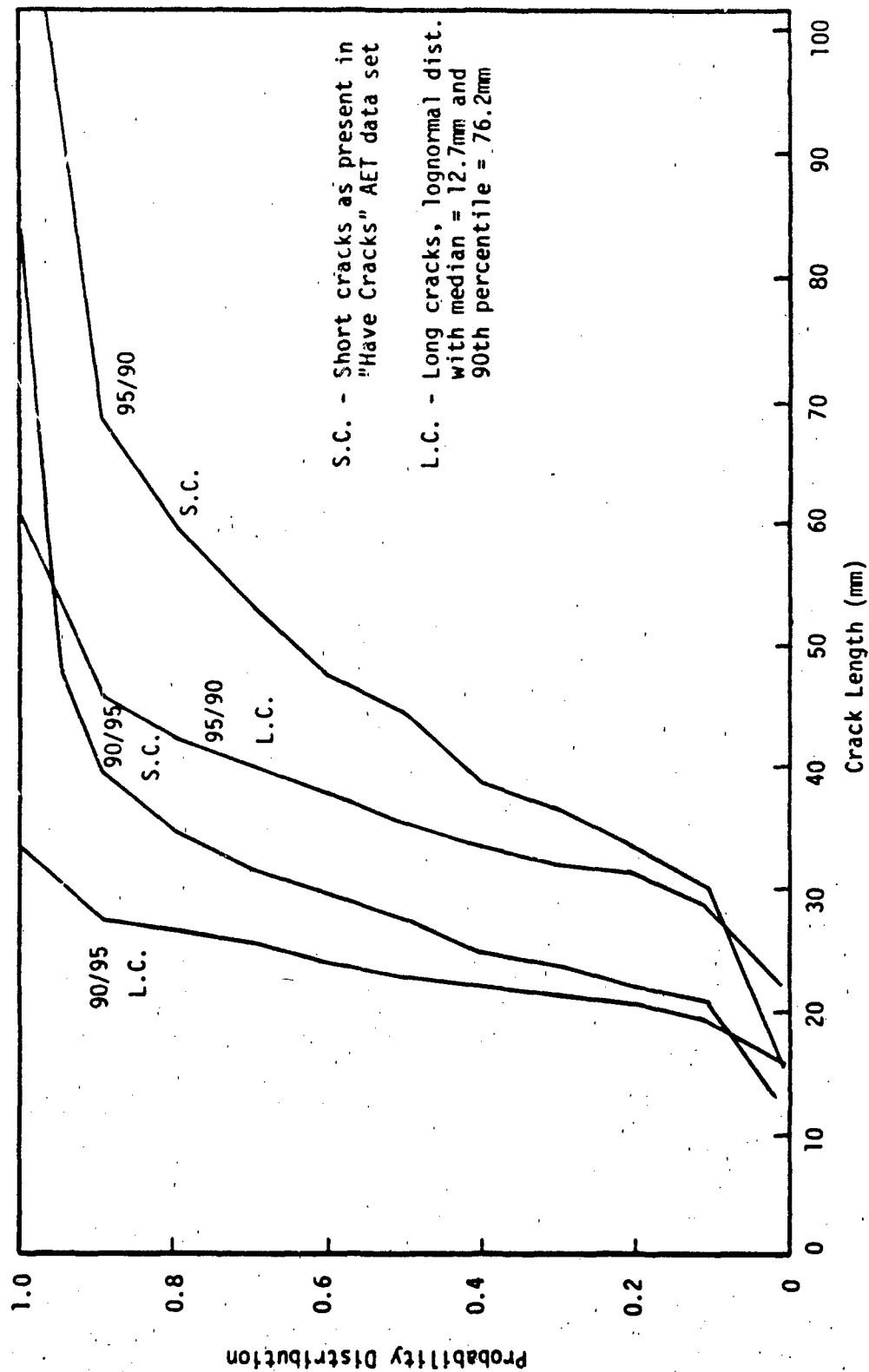


Figure 22. Comparison of Specimen Crack Sizes in an NDE Capability Experiment - AET Environment, Log Odds-Maximum Likelihood Analysis.



In the four cases considered the long crack experiments had less scatter in the estimates of the POD/CL limits than did the short crack experiments. In the log odds-maximum likelihood analysis (Figure 22), the long crack experiments tended to provide estimates closer to the true POD value (i.e. smaller). In the log odds-regression analysis, however, this trend was reversed.

While the effect of specimen crack size distribution has not been sufficiently determined, these comparisons definitely indicate that the sizes of the cracks in a NDE capability demonstration program should be considered as an experimental control to the extent possible.

#### 4.2.4 Comparison of Scatter in Detection Probabilities at a Fixed Crack Length

In an effort to isolate the causes of the large degree of scatter in the estimates of the POD/CL limits, it was postulated that this scatter could be caused by the relatively poor correlation of crack detection with crack length that was present in the "Have Cracks" data. To test this hypothesis, NDE simulations were performed for the AET-long cracks environment but with standard error of deviations about the POD function reduced by a factor of 4. Figure 23 depicts two sets of 95 percent confidence bounds on the detection probabilities from the AET environment. The outside bands would be expected to encompass approximately 95 percent of the individual crack detection probabilities (c.f. Figure 8). The inside bands were generated dividing the "Have Crack" standard error by 4. As can be seen, the detection probabilities for the reduced standard error are much better correlated with crack size.

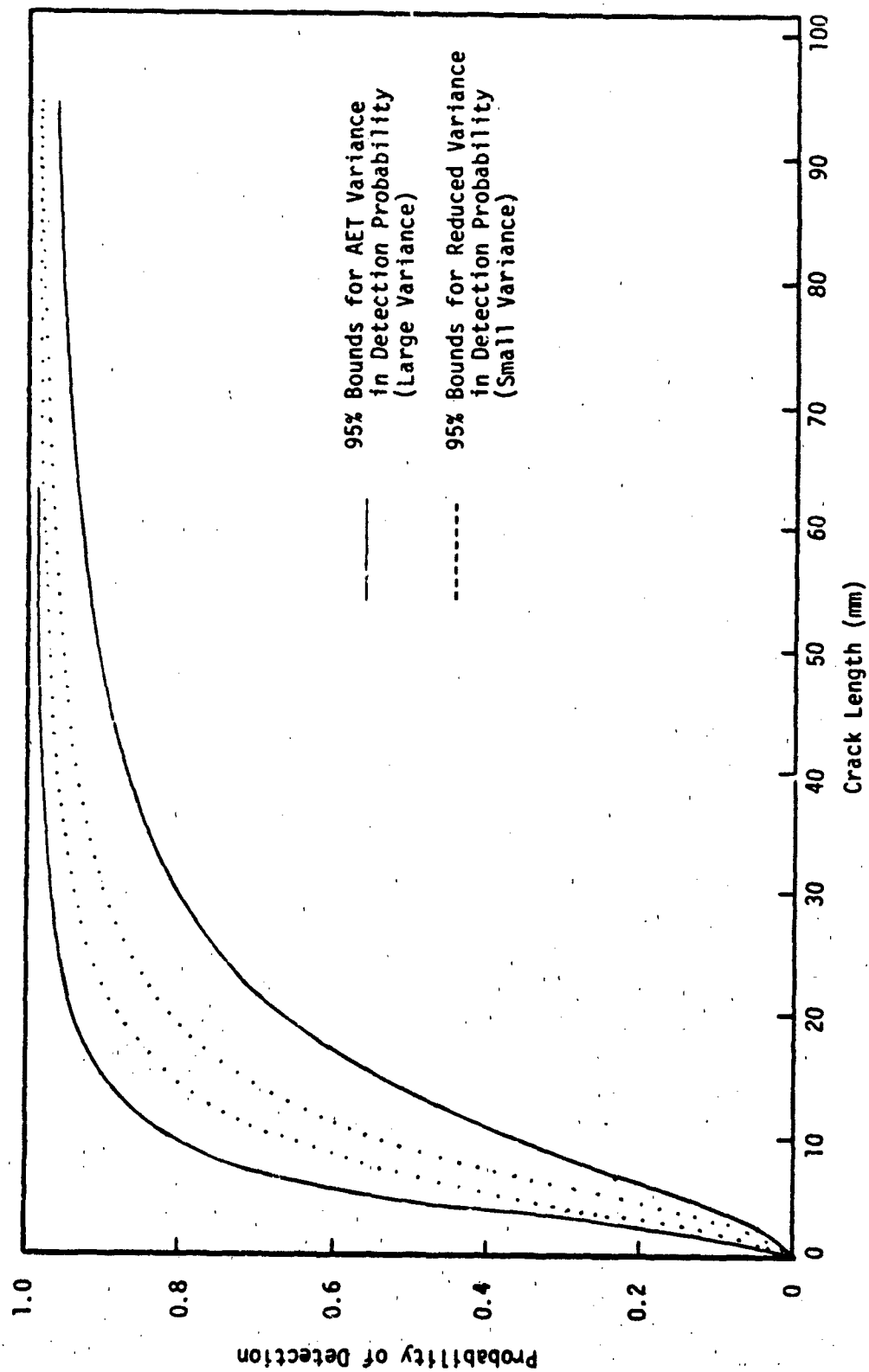


Figure 23. Comparison of Degrees of Scatter in Detection Probabilities.

Figures 31 through 34 present the distributions of 90/95 and 95/90 limits obtained from the simulations for both degrees of variability in the AET-long cracks environment. Increasing the correlation of crack detection probabilities with crack length (i.e., reducing the scatter of individual crack detection probabilities about the POD function) produced two changes. First, smaller POD/CL estimates were obtained from the less scattered detection probabilities even though the true POD was constant in both sets of experiments. Second, the variability of the POD/CL estimates was reduced but not by an amount that would have any real practical significance. Thus, it was concluded that improving the degree of correlation of detection probabilities with crack length will not have a practical effect on the precision of the estimates of the POD/CL limits.

#### 4.3 DISCUSSION

Analysis of the results of the simulated NDE experiments lead to three major conclusions:

- 1) The large degree of variability in the POD/CL crack length estimates for POD values of 0.9 and greater indicates that such estimates are not reproducible if an NDE capability experiment would be repeated.
- 2) Both the magnitude and scatter in the POD/CL estimates are significantly influenced by the crack sizes in the experiment.
- 3) The variability of the POD/CL estimates is not primarily due to the lack of a strong correlation of detection probabilities of individual cracks with crack length.

These conclusions imply that the scatter in the POD/CL estimates is inherent to the analysis procedure and only indirectly to the NDE capability.

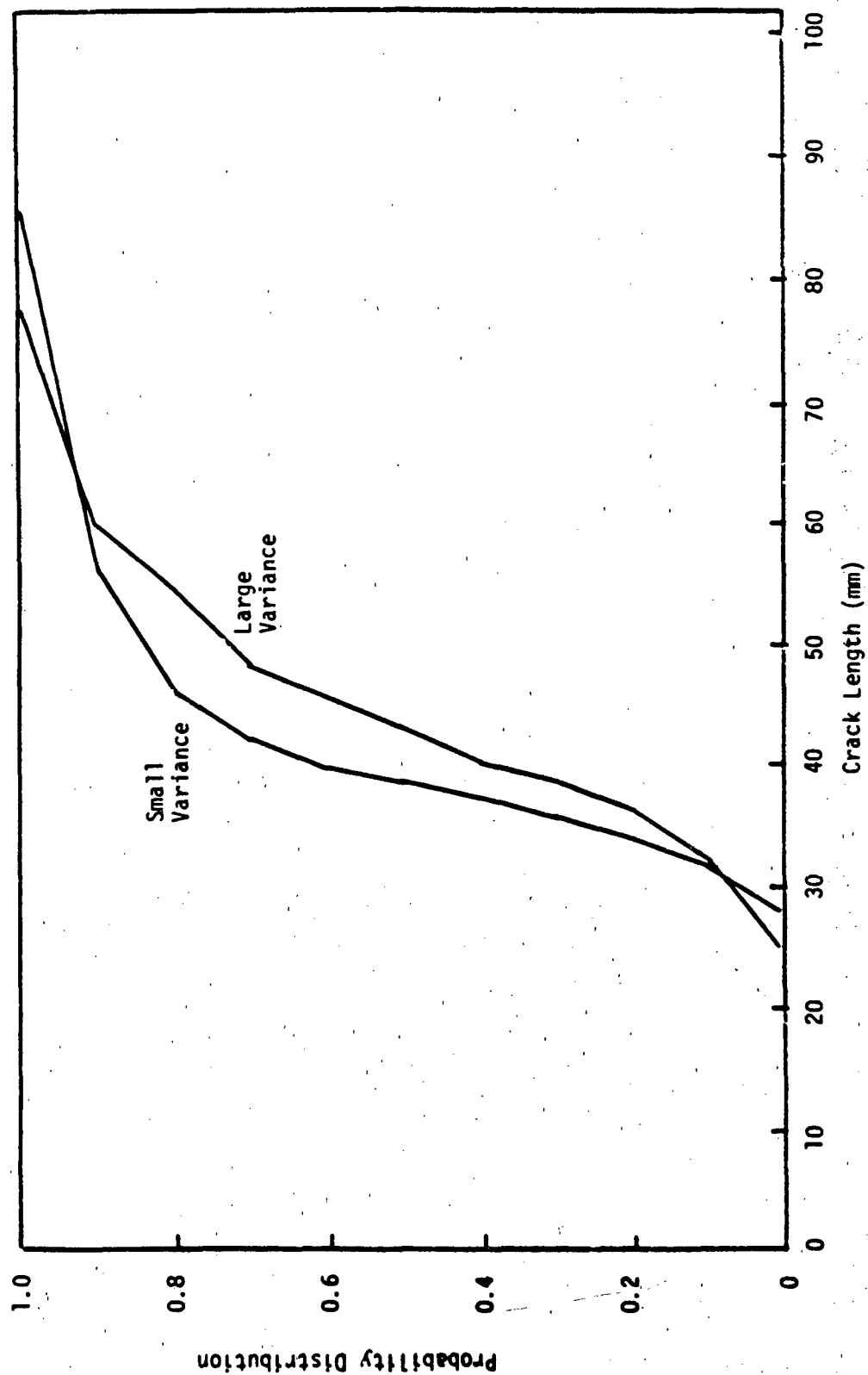


Figure 24. Effect of Scatter in Individual Crack Detection Probabilities on 90/95 Limits, AET-Long Cracks, Log Odds-Regression.

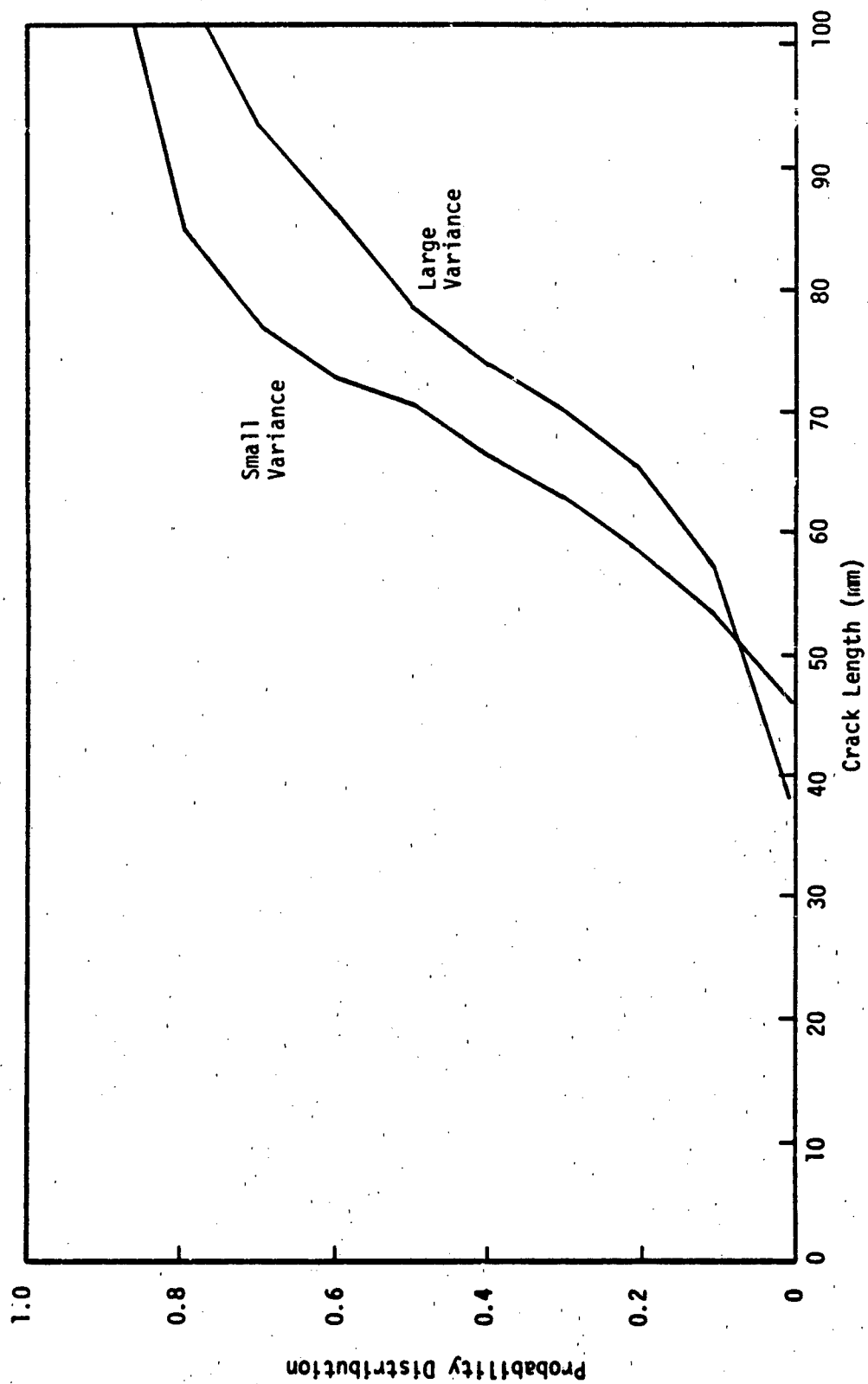


Figure 25. Effect of Scatter in Individual Detection Probabilities on 95/90 Limits, AET - Long Cracks, Log Odds-Regression.

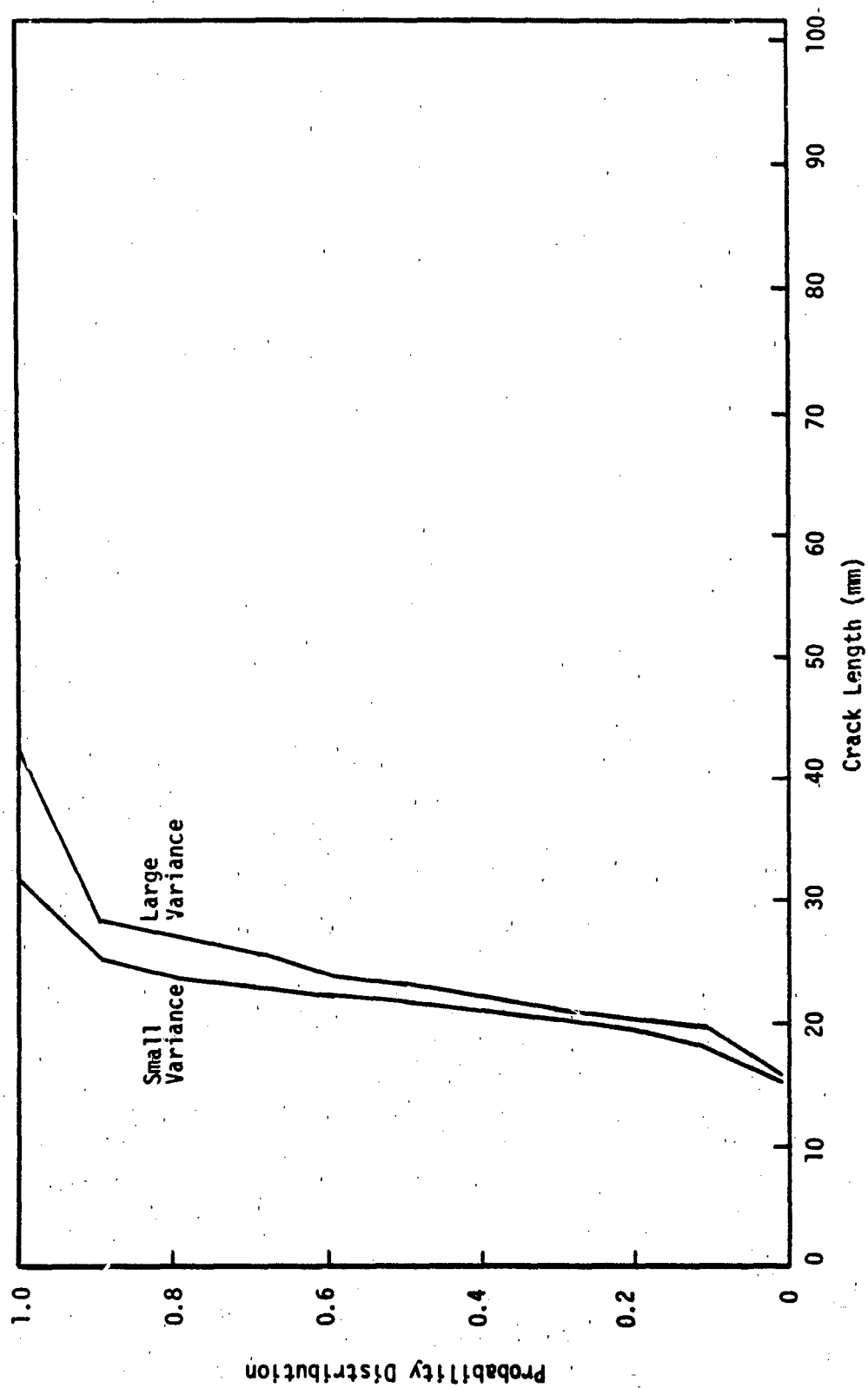


Figure 26. Effect of Scatter in Individual Crack Detection Probabilities on 90/95 Limits, AET - Long Cracks, Log Odds-Maximum Likelihood.

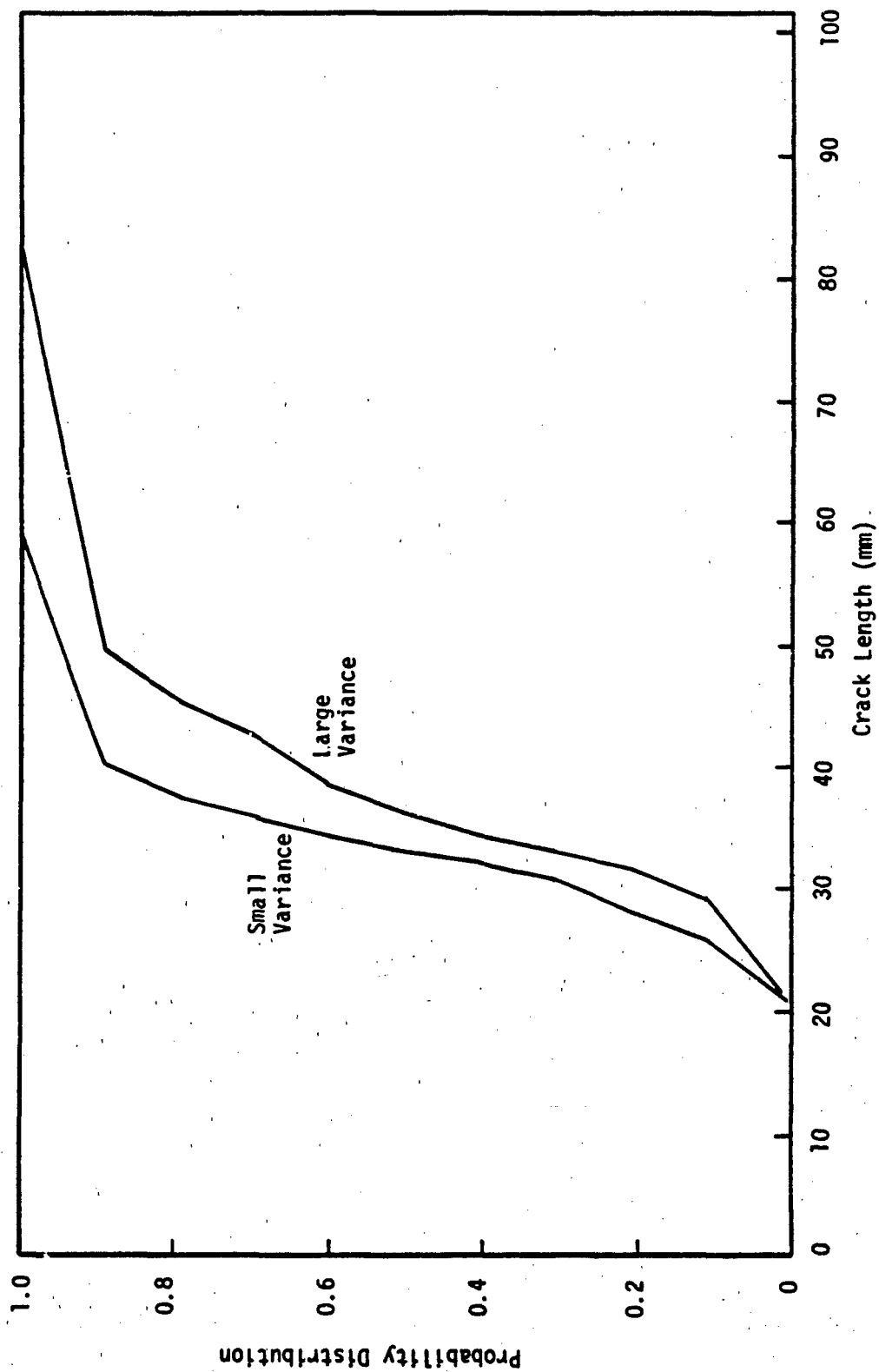


Figure 27. Effect of Scatter in Individual Crack Detection Probabilities on 95/90 Limits, AET, Long Cracks, Log Odds-Maximum Likelihood.

To further explore the instability of crack length estimates corresponding to high POD values, consider the shape of the model for POD as a function of crack length. Available data from NDE reliability experiments indicate that at least some of the longer length cracks fail to be detected. Realistic POD models will account for these misses by asymptotically approaching one. Simple geometric considerations lead to the conclusion that estimates of crack lengths corresponding to POD values in the flat portion of the curve are very sensitive to "errors" in the POD value. See Figure 11. Since the POD value is being estimated statistically, very large sample sizes would be required to reduce the "error" in the POD estimate to yield a precise corresponding crack length.

It is theoretically possible to have an NDE system for which the slope of the POD curve is sufficiently steep, that reasonably precise estimates of the crack length corresponding to a POD of 0.90 or 0.95 can be obtained. Such POD curves have not yet been shown to occur in field applications since human factors as well as inspection hardware influence the capability of the system. Even if such a system were available, however, attempts to characterize it in terms of higher POD levels (say 0.99 or 0.999) would lead to the same lack of precision in the POD/CL estimates.

Assuming the simulations are a reasonable approximation to NDE capability demonstrations, the preceding discussion indicates that POD/CL crack length estimates are quite unstable and, therefore, unacceptable for use as single number crack length characterizations. Actually, the 90/95 limit is commonly quoted by the Air Force because the analysis methods could more often provide this estimate rather than a desire to determine the crack length for which 90 percent of the cracks will be detected. It was generally assumed that the 90/95 limit was "conservative" for a system but no measure of the conservativeness was quoted. Further, this characterization of NDE capability is extremely difficult to use in a meaningful risk analysis.



In the structural maintenance system application, the quantity of real interest is the probability that cracks longer than a fixed value will pass undetected. This probability depends on both the POD distribution and the sizes of the cracks that are present in the structure. In particular, let  $H(a)$  represent the probability of having a crack greater than or equal to  $a$  in the structure and failing to detect it during an inspection whose probability of detection is given by  $POD(a)$ . Then

$$H(a) = \int_a^{\infty} [1 - POD(x)] f(x) dx \quad (26)$$

where  $f(x)$  is the probability density function of the crack sizes in the structure to be inspected.

To illustrate this calculation, Figure 28 presents  $H(a)$  for two crack size distributions and the NDE capability as determined for the AET environment ( $\alpha = -2.9$ ,  $\beta = 1.7$ ). The crack sizes were assumed to have Weibull distributions with shape parameter equal to 2.73. The scale parameters were selected to yield median crack sizes of 12.7mm and 25.4mm. The cumulative distribution of the crack sizes are shown as  $F_1(a)$  and  $F_2(a)$  for the small and large cracks, respectively. The resulting probabilities of having a crack larger than  $a$  and missing it during the inspection are plotted as  $H_1(a)$  and  $H_2(a)$  for the small and large cracks, respectively. The plot for  $H_2(a)$  indicates that in a random inspection from the large cracks distribution, there is a 2 percent chance of having a crack longer than 25mm and not detecting it. On the other hand,  $H_1(a)$  for the small cracks indicates there is practically no chance (within plotting accuracy) of having a crack greater than 25mm and failing to detect. Note that in the small crack distribution there is only a 1 percent chance of having a crack greater than 25mm. Note also that the 90/95 limits for the AET environment were generally greater than 20mm. Therefore, whether or not the 90/95 limit was sufficiently large depends not only on the particular estimate obtained from the NDE experiment but also on the crack sizes in the structure to be inspected.

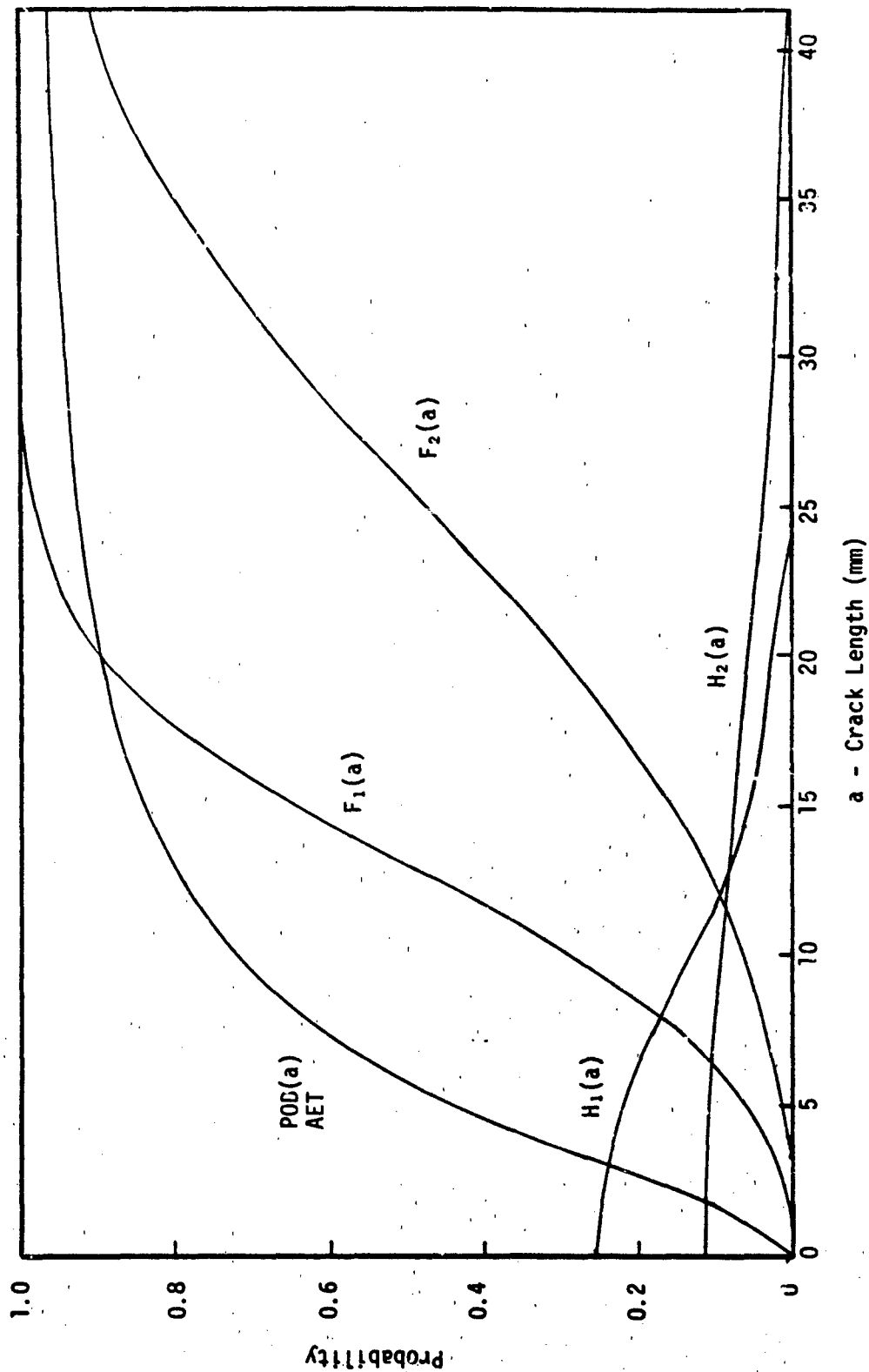


Figure 28. Probability of Failing to Detect a Crack Greater Than Indicated Length.

A second calculation that may prove of interest in structural maintenance plans is the probability of not detecting a crack given that its length is greater than a fixed value,  $a$ . This quantity is given by the expression

$$G(a) = \frac{H(a)}{1 - F(a)} \quad (27)$$

and is calculated quite simply from the cumulative crack size distribution,  $F(a)$ , and  $H(a)$  as given by equation (26).

## SECTION 5

### CONCLUSIONS

This program comprised two distinct phases. The conclusions drawn from each of the phases are presented below.

#### 5.1 POD ANALYSIS FRAMEWORK AND REGRESSION MODEL SELECTION

The objectives of this phase of the study were to formulate and demonstrate an analysis framework for quantifying the results of NDE experiments and to select an appropriate POD model based on the data of the "Have Cracks Will Travel" Program. The following conclusions were reached.

1. All cracks of the same length do not have the same detection probability. The strength of the correlation between detection probability and crack length depends on the NDE system (including inspection environment and human factors), crack geometry, and the structural element being inspected.

2. The probability of detection for cracks of a fixed length in a population of details is the mean (average) of the detection probabilities of the individual cracks.

3. Since the POD as a function of crack length is the curve through the mean values, standard regression techniques of statistics can be used to estimate and place valid confidence limits on this POD function. Care must be taken in the determination of a regression model to ensure not only that the equation "fits" the data but also that the deviations from the predicted average values are normally distributed with equal variance for all crack lengths.

4. The log odds model met the three acceptability criteria when applied to the eight data sets of the "Have Cracks Will Travel" program that had sufficient data for meaningful analysis. Six other models which have found wide use in problems of a similar nature did not perform as well on the same data. Since the log odds model was designed to yield both equal variance and normality of deviations from the equation, this model is conditionally

recommended for regression analysis of POD data. However, due to the limited experience in the regression analysis of POD data, the applicability of the log odds model should be verified whenever possible. If necessary, more appropriate models should be used.

5. A method for estimating the parameters of the log odds model using maximum likelihood was derived for the NDE experiment in which each crack is inspected once. A method for placing confidence bounds on the POD function was also established using asymptotic properties of the maximum likelihood estimates. This maximum likelihood method does not require grouping cracks into ranges of similar length and is distinct from the least squares estimates of a regression analysis.

## 5.2 RESULTS OF SIMULATION STUDIES

In the second phase of the program, NDE demonstration programs were simulated using representative capabilities and scatter in detection probabilities as determined for the "Have Cracks Will Travel" data. The objectives of this phase were to compare analysis methods and capability characterizations as expressed by combinations of POD and confidence level. The following conclusions were reached.

1. Given an acceptable model for the regression function, the regression estimates of NDE capability expressed in terms of a confidence limit on a high probability of detection value (i.e. a POD/CL value) are superior to those derived using binomial distribution theory. The regression estimates are closer to the true POD, exhibit less scatter in the distribution of the estimates, and, contrary to binomial methods, always provide an estimate of the desired limit.

2. The magnitude and scatter of the POD/CL values are significantly influenced by the crack sizes employed in the NDE capability experiment.

3. The degree of scatter of the detection probabilities of individual cracks about the POD function has only a secondary effect on the scatter in the POD/CL estimates.

4. Single number characterizations of NDE capability expressed in terms of a probability of detection and a confidence level (POD/CL) display a degree of scatter (i.e. non-reproducibility) that make these characterizations of limited practical use in the evaluation of NDE systems.

5. A more complex characterization of capability will be required for use in the evaluation of NDE systems.

## REFERENCES

1. Packman, P. F., S. J. Klima, R. L. Davies, Jugal Malpani, J. Moyzis, W. Walker, B. G. W. Yee, and D. P. Johnson, "Reliability of Flaw Detection by Nondestructive Inspection," ASM Metal Handbook, Vol. 11, 8th Edition, Metals Park, Ohio, pp. 214-224, 1976.
2. Yee, B. G. W., F. H. Chang, J. C. Couchman, G. H. Lemon, and P. F. Packman, "Assessment of NDE Reliability Data," NASA CR-134991, National Aeronautics and Space Administration, Lewis Research Center, Cleveland, Ohio, 1976.
3. Lewis, W. H., B. D. Dodd, W. H. Sproat, and J. M. Hamilton, "Reliability of Nondestructive Inspections - Final Report," Report No. SA-ALC/MEE 76-6-38-1, United States Air Force, San Antonio Air Logistics Center, Kelly Air Force Base, Texas, 1978.
4. Natrella, M. G., Experimental Statistics, Handbook 91, National Bureau of Standards, 1963.
5. Cox, D. R., The Analysis of Binary Data, Methuen and Co., LTD, London, 1970.
6. Finney, D. J., Statistical Method in Biological Assay, Hafner Publishing Company, New York, 1964.
7. Anderson, Virgil L. and Robert A. McLean, Design of Experiments, Marcel Dekker, Inc., New York, 1974.

PROCESSING PAGE BLANK-NOT FILMED

APPENDIX A  
"HAVE CRACKS WILL TRAVEL" DATA BASE

RESEARCH PAGE BLANK-NOT FILMED



## APPENDIX A

### "HAVE CRACKS WILL TRAVEL" DATA BASE

Under an Air Force program entitled "Reliability of Non Destructive Inspections," personnel from the Lockheed - Georgia Company transported fatigue damaged structural samples to Air Force bases and depots. The samples were inspected by representative inspectors at each facility using current, standard NDE technology. The results of each inspection along with a large body of concomitant data were stored in a data base which has been commonly identified as the "Have Cracks Will Travel" data. At the conclusion of the inspection phase of the program, the cracks in each structural element were found and measured and these data were also incorporated into the data base.

The "Have Cracks" data base contains the results of approximately 22,000 inspections that were made on 174 cracks. Each crack was inspected by 2 or 3 procedures appropriate to the structure and by as many as 107 different inspectors. Since different NDE capabilities were anticipated for the different NDE methods and structures, the data were partitioned into 13 sets as defined by structure type and inspection method. Table A.1 identifies the 13 data sets and lists the number of cracks and number of inspections of each crack of the data sets. Figures A.1 through A.13 present plots of the detection percentages for each crack in each data set.

There were two objectives in the analysis of these data: 1) to determine a regression model which best fit the data as defined by three criteria, and 2) to determine estimates of the parameters of the model and variability of POD values about the model as representative input to a simulation study. The goodness of fit criteria were: 1) the pattern and magnitude of the individual deviations from the regression curve (the residuals); 2) the equality of variance of the residuals for all crack lengths; and, 3) the normality of the residuals. The first criteria concerns the ability of the model to adequately represent the observations. The last two criteria are necessary assumptions for deter-

mining confidence limits on the POD curve and also provide the distributional framework for performing the simulation studies of this program. It should be noted that since detection probabilities are always between zero and one, candidate regression models required transformations of the observed POD and crack length values. Thus, the goodness of fit and residual analyses were performed in the domain of the transformed values.

The sample type E specimens (data sets EEA and EEH) contain only 6 cracks. Although these data were fit to the various regression models, they were ignored in the determination of the best fit. Similarly, the AUA, BEO, and FEA data sets were judged to contain too few inspections per crack to be permitted to influence model selection. The standard deviation of an estimated percentage due only to sampling error is  $\sqrt{p(1-p)/n}$ . Those data sets which have few inspections per crack would have significantly greater variability in the estimates of individual POD values than those with 50 or more inspections per crack. The CUT and FUT sets with 32 and 27 inspections per crack, respectively, were not eliminated as these were considered the transitional sample sizes. Thus, the selection of the POD model was based on 8 of the 13 data sets.

Also included in Figure A.1 through A.13 are the POD curves derived from the log odds model which demonstrate the goodness of fit of this model. Data sets AET, AUT, and BET were selected to represent the "typical" inspection capability for the simulation phase of the study. These choices represent three reasonable degrees of NDE capability and two degrees of scatter of the individual detection probabilities about the mean POD function.

TABLE A.1  
"HAVE CRACKS WILL TRAVEL" DATA SETS

DATA SET NAME	SAMPLE TYPE	NDI METHOD	NUMBER OF CRACKS	INSPECTIONS PER CRACK
AET	A	Eddy Current - Surface	41	62
AUA	A	Ultrasonic - Automatic	41	4
AUT	A	Ultrasonic	41	54
BEO	B	Eddy Current - Overhead	52	9
BET	B	Eddy Current - Surface	52	94
BRT	B	X-Ray	52	59
CPT	C	Penetrant	41	63
CUT	C	Ultrasonic - Manual	41	32
EEA	E	Eddy Current Bolt Hole - Automatic	6	21
EEH	E	Eddy Current Bolt Hole - Manual	6	107
FEA	F	Eddy Current Bolt Hole - Automatic	34	13
FEH	F	Eddy Current Bolt Hole - Manual	34	79
FUT	F	Ultrasonic	34	27

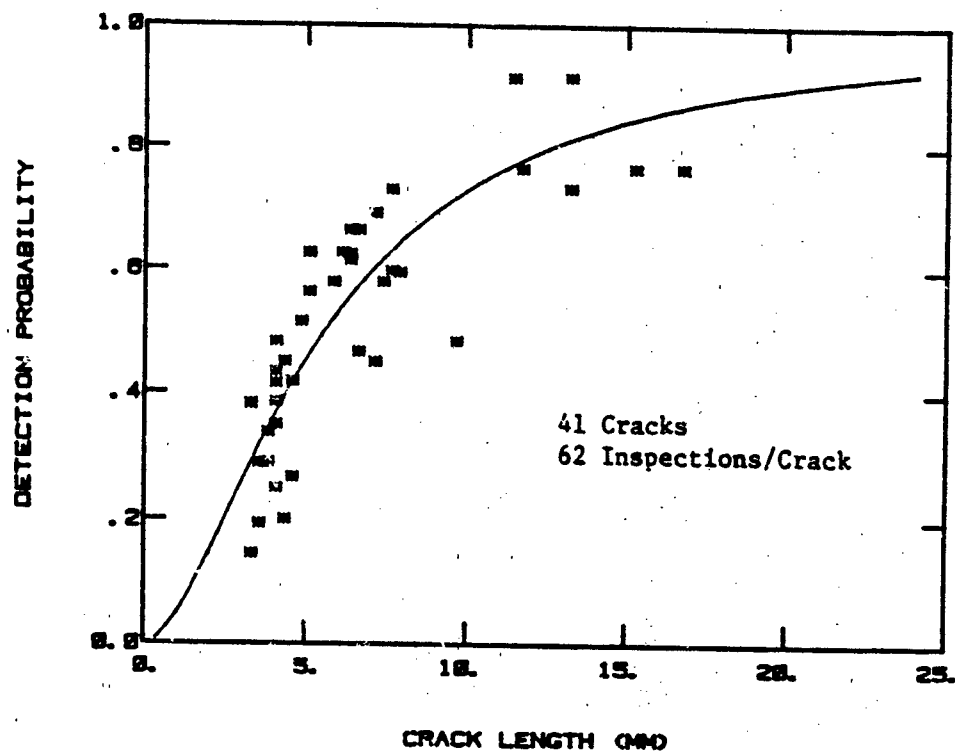


Figure A.1 Detection percentages and mean log odds model for cracks of AET data set - Have Cracks Will Travel Data Base.

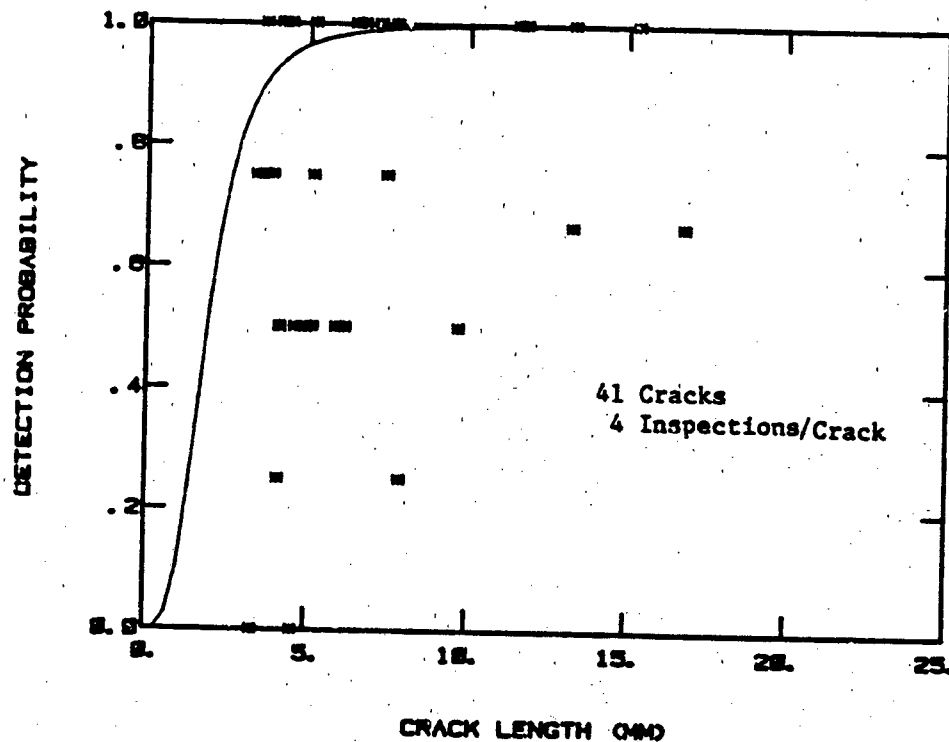


Figure A.2 Detection percentages and mean for log odds model for cracks of AUA data set - Have Cracks Will Travel Data Base.

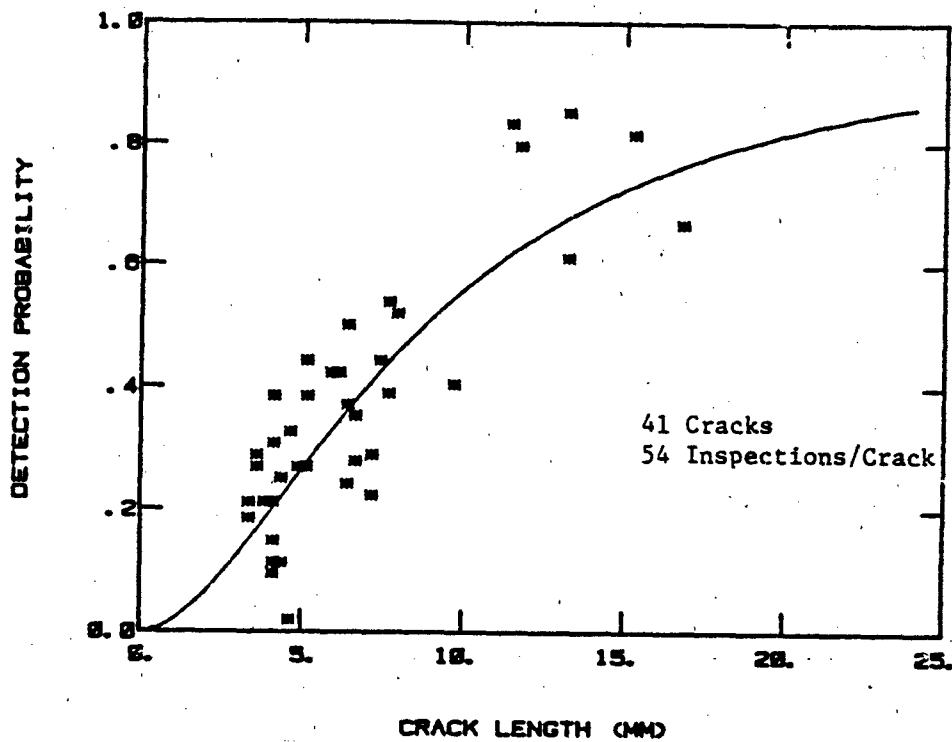


Figure A.3 Detection percentages and mean log odds model for cracks of AUT data set - Have Cracks Will Travel Data Base.

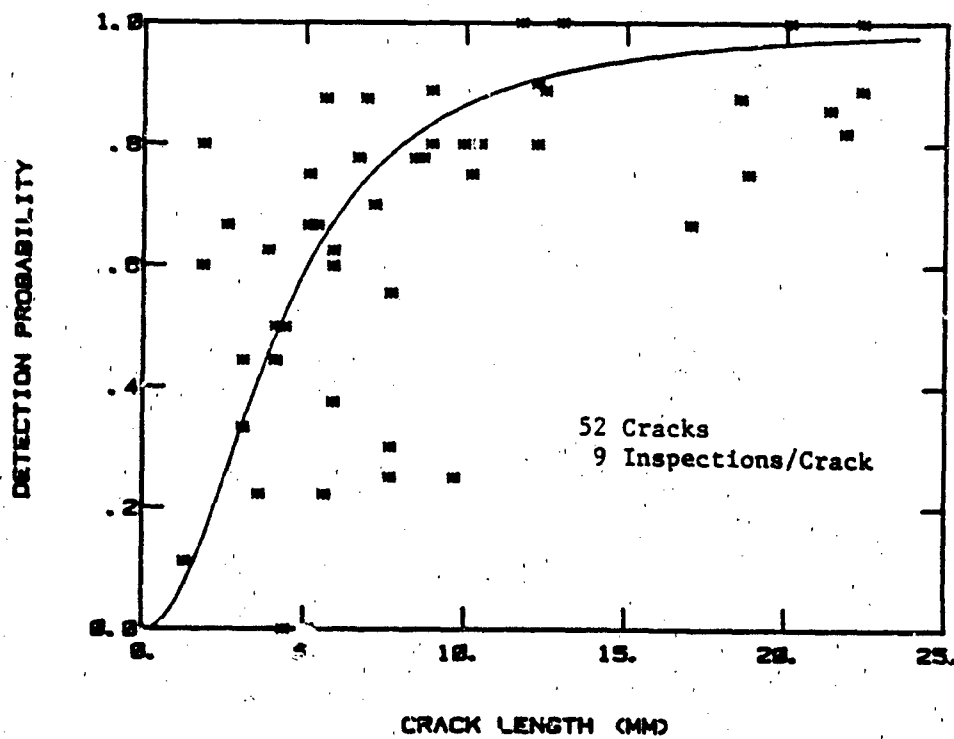


Figure A.4 Detection percentages and mean log odds model for cracks of BEO data set - Have Cracks Will Travel Data Base.

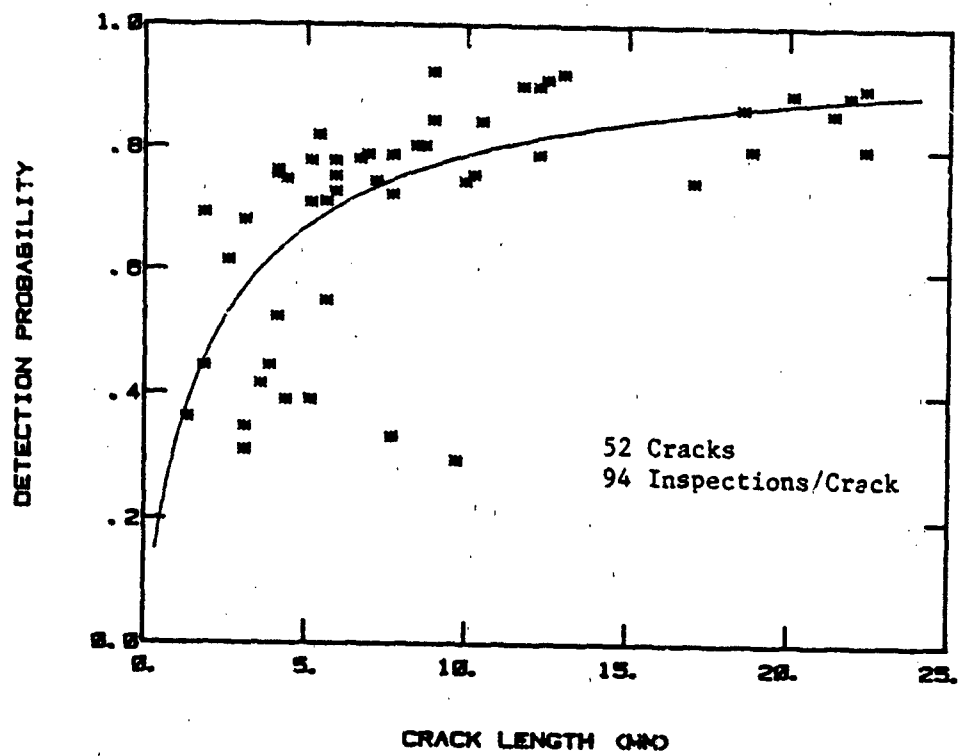


Figure A.5 Detection percentages and mean log odds model for cracks of BET data set - Have Cracks Will Travel Data Base.

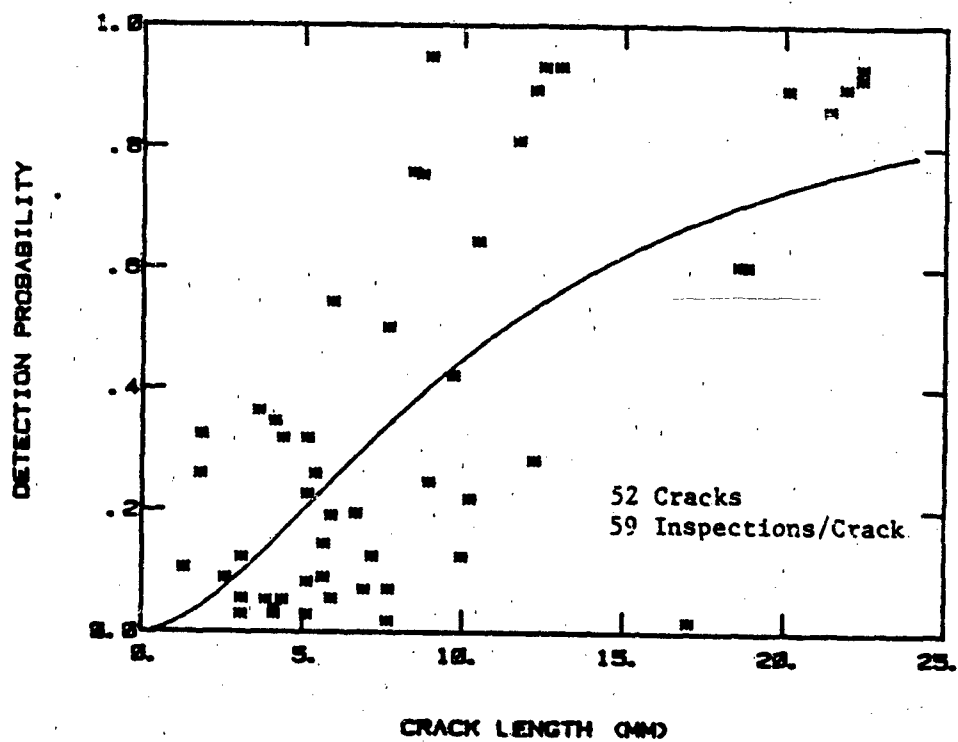


Figure A.6 Detection percentages and mean log odds model for cracks of BRT data set - Have Cracks Will Travel Data Base.

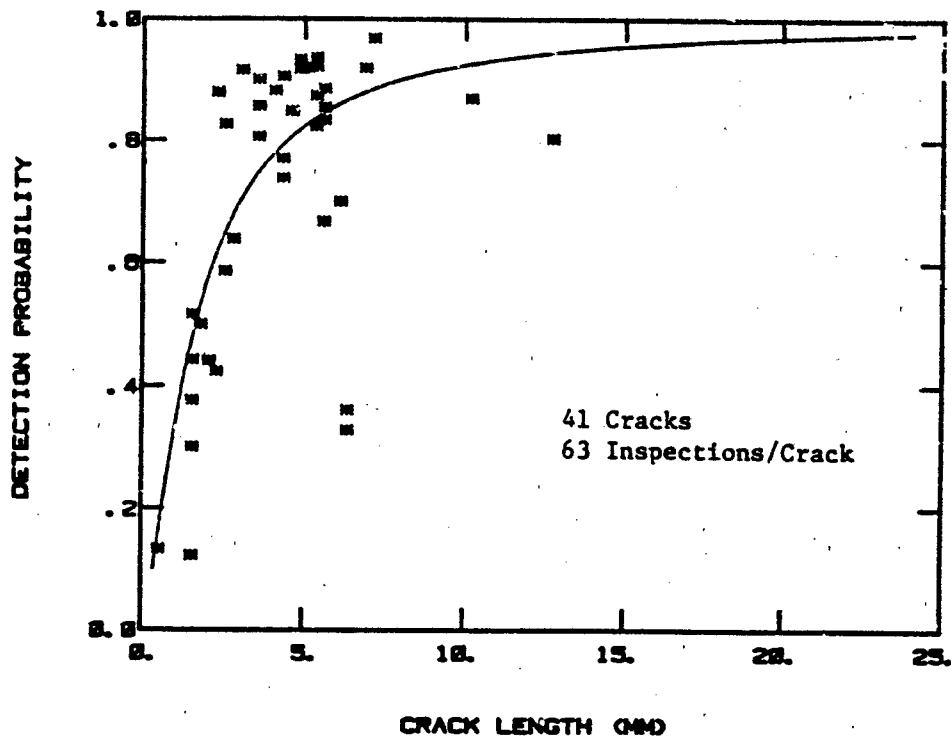


Figure A.7 Detection percentages and mean log odds model for cracks of CPT data set - Have Cracks Will Travel Data Base.

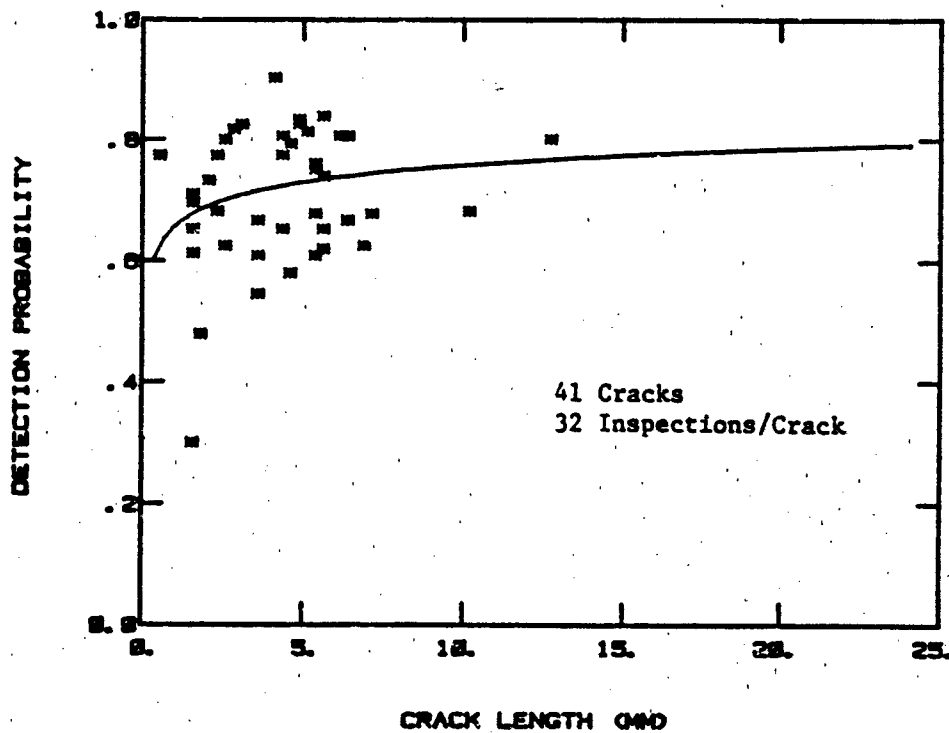


Figure A.8 Detection percentages and mean log odds model for cracks of CUT data set - Have Cracks Will Travel Data Base.

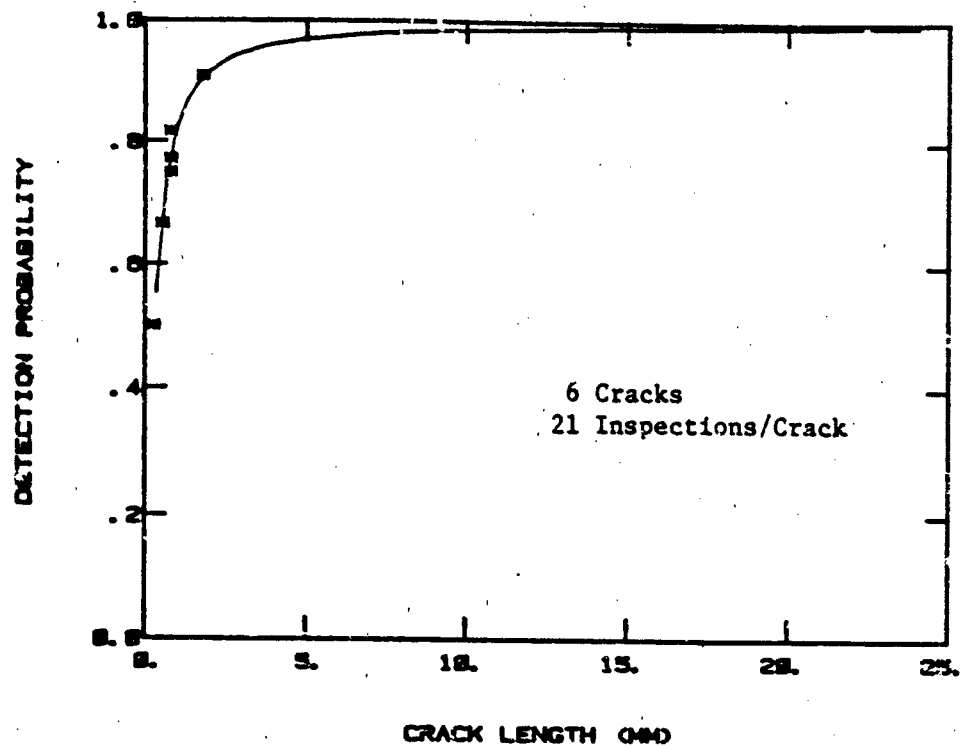


Figure A.9 Detection percentages and mean log odds model for cracks of EEA data set - Have Cracks Will Travel Data Base.

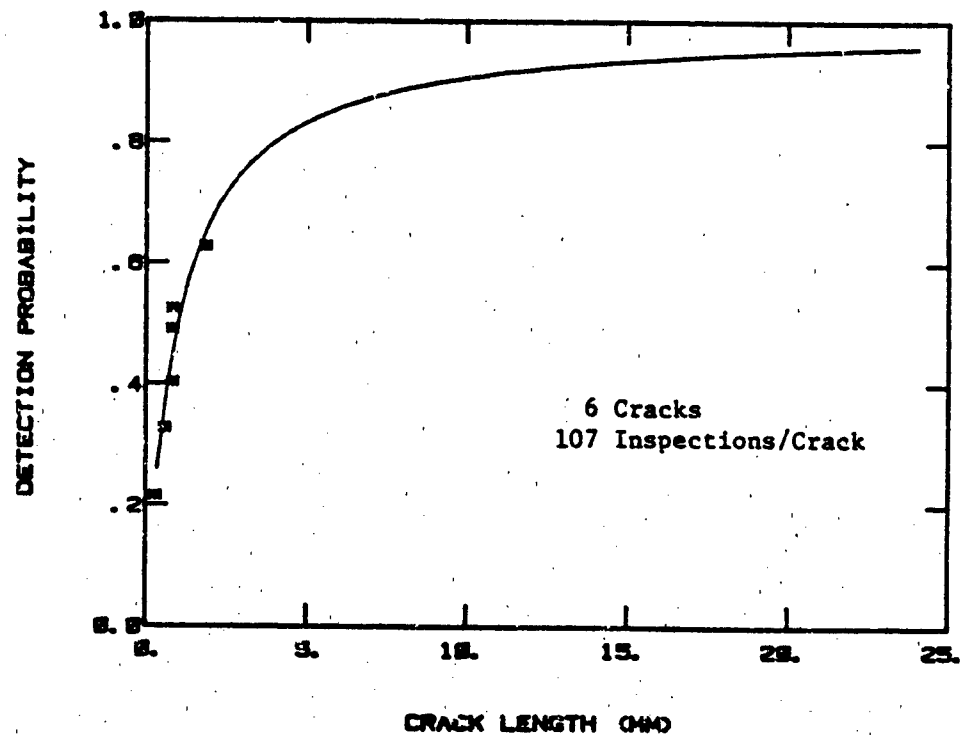


Figure A.10 Detection percentages and mean log odds model for cracks of EEH data set - Have Cracks Will Travel Data Base.



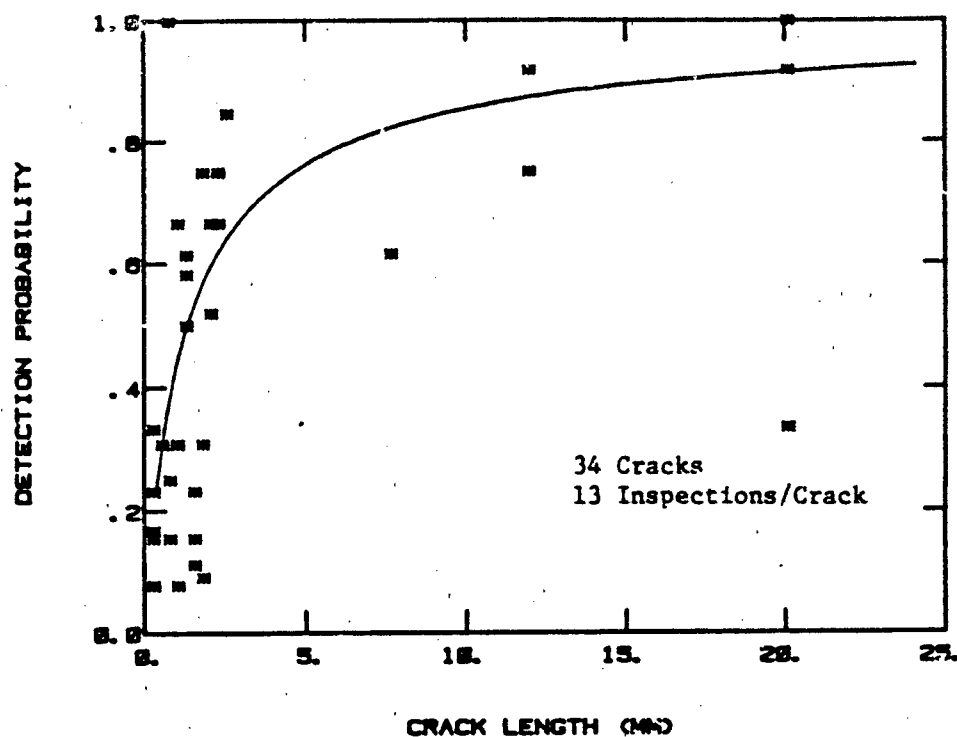


Figure A.11 Detection percentages and mean log odds model for cracks of FEA data set - Have Cracks Will Travel Data Base.

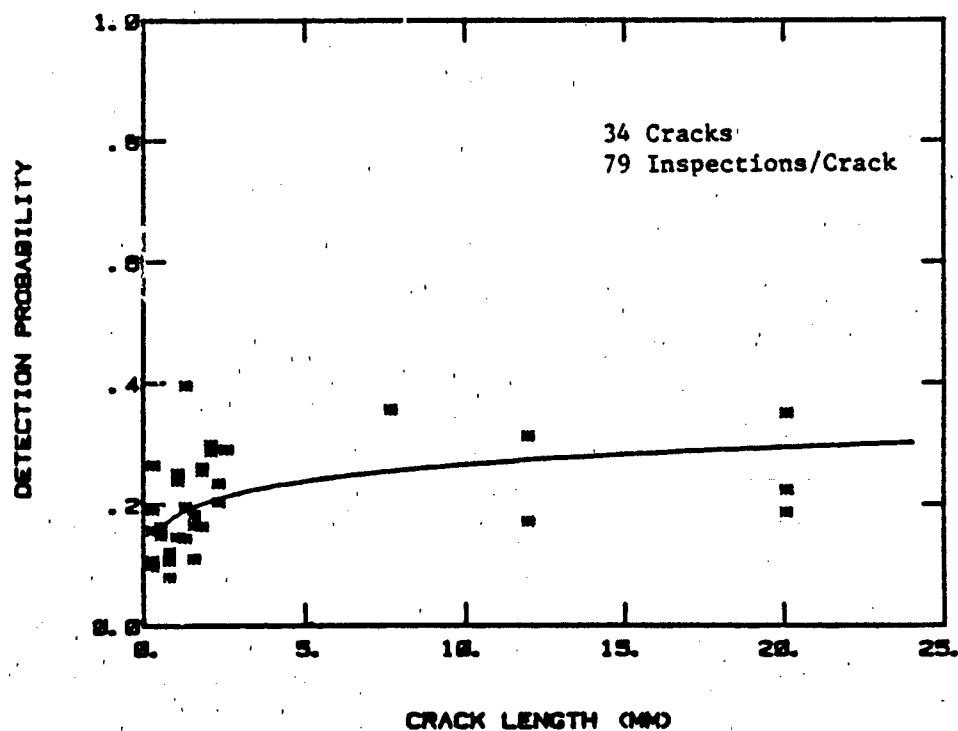


Figure A.12 Detection percentages and mean log odds model for cracks of FEH data set - Have Cracks Will Travel Data Base.

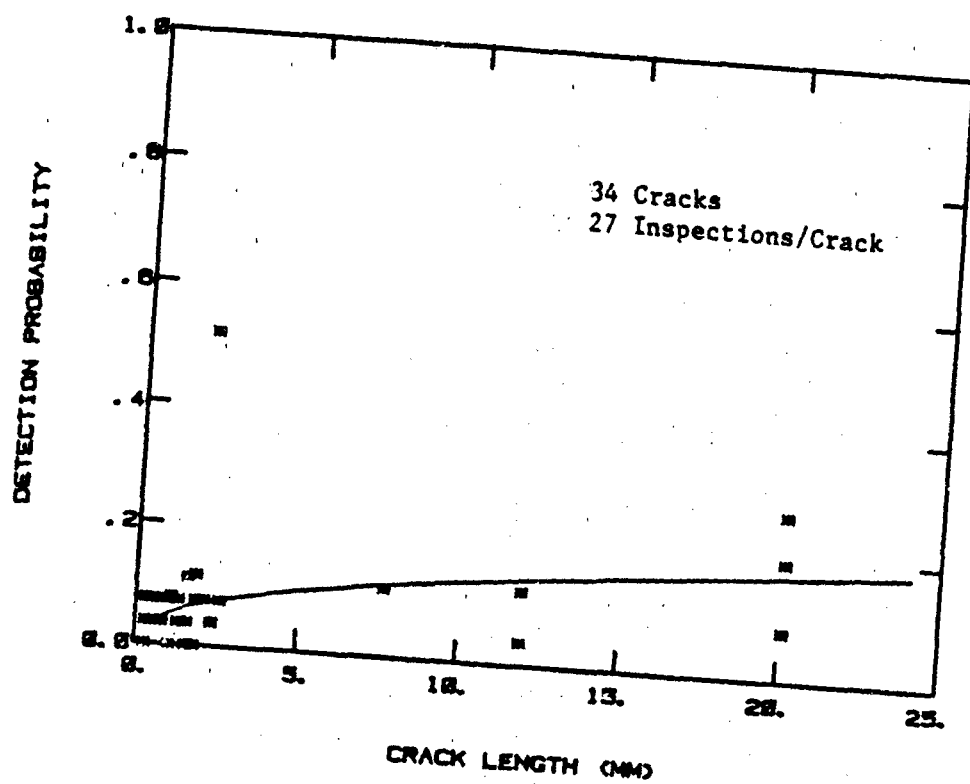


Figure A.13 Detection percentages and mean log odds model for cracks of FUT data set - Have Cracks Will Travel Data Base.